Envelope Models for Parsimonious and Efficient Multivariate Linear Regression

R. Dennis Cook[1], Bing Li[2] and Francesca Chiaromonte[2]

[1]University of Minnesota and [2]Pennsylvania State University

May 21, 2009

Abstract

We propose a new parsimonious version of the classical multivariate normal linear model, yielding a maximum likelihood estimator (MLE) that is asymptotically less variable than the MLE based on the usual model. Our approach is based on the construction of a link between the mean function and the covariance matrix, using the minimal reducing subspace of the latter that accommodates the former. This leads to a multivariate regression model, which we call the *envelope model*, where the number of parameters is maximally reduced. The MLE from the envelope model can be *substantially* less variable than the usual MLE, especially when the mean function varies in directions that are orthogonal to the directions of maximum variation for the covariance matrix.

Key words and phrases: Discriminant analysis, Functional data analysis, Grassmann manifolds, Invariant subspaces, Principal components, Reduced rank regression, Reducing subspaces, Sufficient dimension reduction.

1 Introduction

A cornerstone of multivariate analysis is the following multivariate linear regression model

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the random response vector, $\mathbf{X} \in \mathbb{R}^p$ is a non-stochastic vector of predictors and the error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is normally distributed with mean $\mathbf{0}$ and unknown covariance matrix $\boldsymbol{\Sigma} \geq \mathbf{0}$ (see Christensen, 2001, for background). If $\mathbf{X}$ is random during sampling then the model is conditional on the observed values of $\mathbf{X}$. This conditioning, which is common practice in regression, was discussed by Aldrich (2005) from an historical perspective. The intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$ is an unknown parameter vector and $\boldsymbol{\beta}$ is an unknown parameter matrix of dimensions $r \times p$. Model (1) has a total of $r+pr+r(r+1)/2$ unknown real parameters when $\boldsymbol{\Sigma} > \mathbf{0}$, and it may be a rather coarse tool if this number is large. Variations have been developed to sharpen its abilities. Notable among them is the class of reduced-rank regressions, which allow for the possibility that $\text{rank}(\boldsymbol{\beta}) < \min(p, r)$ (Reinsel and Velu, 1998). In this article we propose a new version of model (1) that yields a maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ with the potential to be substantially less variable asymptotically than the usual MLE. In the remainder of this section we discuss our motivation and describe its implications informally, outline the rest of the article and establish notation for the technical developments that begin in Section 2.

1.1 Motivation

Our primary motivation comes from the simple observation that some characteristics of the response vector could be unaffected by changes in the predictors. Multiple responses are incorporated in many regressions in an effort to encapsulate changes in the distribution of an experimental or sampling unit as the predictors vary. For example, several

anatomical measurements might be taken on individual skulls to compare populations, milk production might be measured on dairy cows at several points during the lactation cycle, hematological measures might be taken on patients at several times following a drug treatment or spectral readings might be taken on samples at several wavelengths. In the same vein, multiple distances and angular measurements are used to model human motion in ergonomic studies (e.g. Faraway and Reed, 2007), and multiple biomarkers are used as responses when studying dietary patterns that affect coronary artery disease (Hofmann, Zyriax, Boeing and Windler, 2004). In these types of multivariate regressions it may be reasonable to allow for the possibility that aspects of the response vector are stochastically constant as the predictors vary.

Assuming model (1), suppose that we can find an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r \times r}$ that satisfies the following two conditions: (i) $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\boldsymbol{\Gamma})$, and (ii) $\boldsymbol{\Gamma}^T\mathbf{Y}$ is conditionally independent of $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ given $\mathbf{X}$. Condition (i) is not restrictive by itself, since at least one, and typically infinitely many semi-orthogonal matrices $\boldsymbol{\Gamma}$ exist with a span containing $\text{span}(\boldsymbol{\beta})$. Under this condition the marginal distribution of $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ does not depend on $\mathbf{X}$. However, $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ may still provide information about the regression through its association with $\boldsymbol{\Gamma}^T\mathbf{Y}$. This possibility is ruled out by condition (ii). Together, conditions (i) and (ii) imply that $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ is marginally independent of $\mathbf{X}$ *and* conditionally independent of $\mathbf{X}$ given $\boldsymbol{\Gamma}^T\mathbf{Y}$. If $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ were known the analysis could be facilitated by using the transformed response $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)^T\mathbf{Y}$, and then backtransforming to the original scale after estimation. In practice we will not normally know a suitable transformation; nevertheless the possibility that such a transformation may exist has important implications for the analysis. In this setting it can be verified that

$$\boldsymbol{\Sigma} = \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}}, \tag{2}$$

where $\mathbf{P}_{\boldsymbol{\Gamma}}$ is the projection onto $\text{span}(\boldsymbol{\Gamma})$ in the usual inner product and $\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{I}_r - \mathbf{P}_{\boldsymbol{\Gamma}}$.

More precisely, given condition (i), condition (ii) is equivalent to equality (2). The crucial point here is that conditions (i) and (2) establish a *parametric link* between $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that is the key for the new methodology proposed in this article. However, this link is not now well defined because there may still be infinitely many subspaces span($\boldsymbol{\Gamma}$) that satisfy the conditions. Section 2 is devoted to algebraic background necessary to construct the unique smallest subspace span($\boldsymbol{\Gamma}$) that satisfies (2) and contains span($\boldsymbol{\beta}$). This minimal subspace, which we call the $\boldsymbol{\Sigma}$-*envelope of* span($\boldsymbol{\beta}$) in full, and the *envelope* for brevity, is then used as a parameter in the *envelope model for multivariate linear regression* defined in Section 3. For now we proceed as if span($\boldsymbol{\Gamma}$) were the envelope.

The full space $\mathbb{R}^r = \text{span}(\mathbf{I}_r)$ trivially contains span($\boldsymbol{\beta}$) and satisfies decomposition (2). If $\mathbb{R}^r$ is the envelope, then the entire response vector $\mathbf{Y}$ is relevant to the regression, a finding that could be useful in its own right. We expect $\mathbb{R}^r$ to be the envelope when r is small and the responses are carefully chosen to reflect distinct aspects of the sampling units. However, we also expect that redundant or irrelevant information will be present in the kinds of applications we have in mind, particularly when many responses are measured in an effort to capture characteristics of the sampling units that vary with the predictors.

Instances of this may occur as a consequence of reasoning about underlying processes. This is the case, for example, in the context of large-scale gene expression data from microarrays. Our argument is tantamount to that used by Leek and Storey (2007) when proposing their method of surrogate variable analysis. Suppose we would like to regress a vector $\mathbf{Y}$ of many (perhaps thousands) gene expression readings on a set of covariates $\mathbf{C}$ (these may comprise environmental factors, treatments or clinical outcomes). Assume that there is an "ideal" vector $\boldsymbol{\nu} \in \mathbb{R}^d$ of latent variables connecting these covariates and the expression levels, so that $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\epsilon}_0$ – where $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix and $\text{var}(\boldsymbol{\epsilon}_0) = \sigma^2\mathbf{I}_r$, as argued by Leek and Storey. Since $\boldsymbol{\nu}$ is unobserved, we write $\boldsymbol{\nu} =$

$\mathrm{E}(\boldsymbol{\nu}|\mathbf{C})+\boldsymbol{\epsilon}$ and then substitute into the model to obtain $\mathbf{Y} = \boldsymbol{\mu}+\boldsymbol{\Gamma}\mathrm{E}(\boldsymbol{\nu}|\mathbf{C})+\boldsymbol{\Gamma}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_0$. The covariates $\mathbf{C}$ might provide only partial information on $\boldsymbol{\nu}$, so some coordinates of $\mathrm{E}(\boldsymbol{\nu}|\mathbf{C})$ could be constant, with the consequence that $\mathrm{E}(\boldsymbol{\nu}|\mathbf{C})$ varies in fewer than d dimensions. The modeling process can be viewed as providing a representation for the unknown conditional mean $\mathrm{E}(\boldsymbol{\nu}|\mathbf{C}) = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}\mathbf{X}(\mathbf{C})$, where $\mathbf{X}$ is the vector of predictors included in the model. As represented, $\mathbf{X}$ is a function of $\mathbf{C}$ and might contain transformations of the measured covariates or interactions among them. Assuming that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\epsilon}_0$, this then leads to the multivariate linear model (1) with $\boldsymbol{\alpha} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\gamma}_0$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\gamma}$, $\boldsymbol{\varepsilon} = \boldsymbol{\Gamma}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_0$, and

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\mathrm{var}(\boldsymbol{\epsilon})\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_r \\ &= \boldsymbol{\Gamma}(\mathrm{var}(\boldsymbol{\epsilon}) + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{3}$$

Since $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$ we have an instance of (2) with $\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}(\mathrm{var}(\boldsymbol{\epsilon}) + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T$ and $\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}} = \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$. The same essential reasoning can be applied in the context of multivariate calibration, where $\mathbf{Y}$ is the vector of spectral readings and $\boldsymbol{\nu}$ depends on the concentrations of interest and all other characteristics of the sample that affect the readings.

Decomposition (2) implies that the eigenvectors of $\boldsymbol{\Sigma}$ fall in either the envelope $\mathrm{span}(\boldsymbol{\Gamma})$ or its orthogonal complement $\mathrm{span}(\boldsymbol{\Gamma}_0)$. The corresponding eigenvalues of $\boldsymbol{\Sigma}$ need not be partitioned in any particular order, since (2) does not presume any relation between the magnitudes of the two terms comprising $\boldsymbol{\Sigma}$. The greatest gains in efficiency will occur when the first term on the right of (2), i.e. $\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}}$, is associated with the smaller eigenvalues of $\boldsymbol{\Sigma}$. However, efficiency gains can also occur under (3), where the envelope captures the leading eigenvectors of $\boldsymbol{\Sigma}$. Relatedly, the estimated error covariance matrix $\widehat{\boldsymbol{\Sigma}}$ for these regressions often contains a few large eigenvalues followed by a large "tail space" of relatively small eigenvalues of similar size. One can think of this

as the sample counterpart of a population error variability structure with a few leading directions, and a large tail space of approximately spherical spread. This structure is a useful descriptor not just for microarray data, but also for other large-scale genomic data; we recently described it for frequencies of short alignment patterns in a comparative genomic study of regulatory elements (sections of nuclear DNA that determine the activation of genes; Cook, Li and Chiaromonte, 2007, Fig. 2).

The connection with the eigenstructure of $\mathbf{\Sigma}$ can be used to provide some intuition about the mechanisms that produce efficiency gains in our approach. Consider a regression in which $p = 1$, and $\mathbf{\Sigma} > \mathbf{0}$ is known and has distinct eigenvalues. Knowledge of $\mathbf{\Sigma}$ alone does not alter the MLE of $\boldsymbol{\beta}$. However, if we also know that $\boldsymbol{\beta}$ falls in the span of, say, the last eigenvector $\mathbf{v}_r$ of $\mathbf{\Sigma}$, then span($\mathbf{v}_r$) is the envelope and we can use a simple univariate linear regression model with response $\mathbf{v}_r^T\mathbf{Y}$ to estimate the direction and length of $\boldsymbol{\beta}$. If the eigenvalue of $\mathbf{\Sigma}$ corresponding to $\mathbf{v}_r$ is substantially smaller than the largest eigenvalue, then the MLE based on $\mathbf{v}_r^T\mathbf{Y}$ will have substantially smaller variation than the usual MLE. Gains can also be realized when $\mathbf{\Sigma}$ is unknown, but we can infer that the envelope is contained in a subspace spanned by a proper subset of the eigenvectors of $\mathbf{\Sigma}$. In full generality, our envelope models are not limited to regressions with $p = 1$, and do not constrain the rank of $\boldsymbol{\beta}$. They do not require $\mathbf{\Sigma}$ to have distinct eigenvalues, or even to be positive definite. However, to focus on the main ideas, we assume throughout this article that $\mathbf{\Sigma} > 0$.

Next, we use a data example to demonstrate the efficiency gains that are possible with our approach. Consider data on $r = 6$ responses, the logarithms of near infrared reflectance at six wavelengths across the range 1680-2310 nm, measured on samples from two populations of ground wheat with low and high protein content (24 and 26 samples, respectively). The mean difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ corresponds to the parameter vector $\boldsymbol{\beta}$ in model (1), with $\mathbf{X}$ representing a binary indicator; $\mathbf{X} = 0$ for high protein wheat, and

$\mathbf{X} = 1$ for low protein wheat. For these data, the standard errors of the six estimated mean differences based on the usual normal-theory analysis under (1) range between 6.4 and 65.8 times the standard errors of the corresponding estimates based on the envelope model. In other words, to achieve comparable standard errors, normal-theory estimates might have to use as many as $65^2 \times 50$ samples where envelope estimates use 50. This example is revisited in Section 7.2.

Reducing redundancy in large data sets has become paramount in an era of high-throughput technologies and fast computing. In many applications, costs are accrued when increasing the number of units, while hundreds or thousands of variables can be recorded on each unit at relatively low expense – which is often done without articulating a specific design at the outset. The resulting data may contain a considerable amount of information that is either irrelevant or redundant for a given purpose. Contemporary statistical theories and methodologies are quickly evolving to adapt to this new reality, with rapid advances in areas such as dimension reduction, sparse variable selection via regularization, and "large-p-small-n" hypotheses testing. The envelope model we introduce uses the error variability structure to create a minimal enclosing of the mean signal in mutivariate data. If these constraints correspond to physical mechanisms, enveloping is a natural way to reflect them; if not, it can still be used as a means of regularization. In either case, controlling the dimension of the envelope can achieve a degree of "eigen sparsity" for the first two moments – arguably the most important descriptors for a broad range of data analyses.

1.2 Outline

Envelopes, which arise from the concepts of invariant and reducing subspaces, are introduced in Section 2. The results in this section, although technical in nature, are immediately relevant to the core developments of this paper. Envelope models for multi-

variate linear regression are described in Section 3, and maximum likelihood estimation of their parameters is developed in Section 4. Selected asymptotic results are presented in Section 5, and a discussion to aid their interpretation is given in Section 6. Section 7 contains simulation and data analysis results. The envelope theory and methods described in Sections 3–7 make use of the error covariance matrix associated with model (1), i.e. the intra-population covariance matrix $\mathbf{\Sigma} = \mathrm{var}(\mathbf{Y}|\mathbf{X})$. They do not involve the marginal covariances $\mathbf{\Sigma_Y} = \mathrm{var}(\mathbf{Y})$ and $\mathbf{\Sigma_X} = \mathrm{var}(\mathbf{X})$. In Section 3.2 we consider some connections among envelopes based on different matrices, and in Section 8 we discuss other contexts in which envelopes might be useful, including reduced rank multivariate models, discriminant analysis, sufficient dimension reduction and some multivariate methods that involve either $\mathbf{\Sigma_Y}$ or $\mathbf{\Sigma_X}$. Section 9 contains some concluding remarks. An on-line supplement to this article with proofs and other technical details is available at http://www.stat.sinica.edu.tw/statistica.

1.3 Notation and definitions

The following notation and basic definitions will be used repeatedly in our exposition. For positive integers r and p, $\mathbb{R}^{r\times p}$ stands for the class of all matrices of dimension $r \times p$, and $\mathbb{S}^{r\times r}$ denotes the class of all symmetric $r \times r$ matrices. For $\mathbf{A} \in \mathbb{R}^{r\times r}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^r$, $\mathbf{A}\mathcal{S} \equiv \{\mathbf{Ax} : \mathbf{x} \in \mathcal{S}\}$. For $\mathbf{B} \in \mathbb{R}^{r\times p}$, $\mathrm{span}(\mathbf{B})$ denotes the subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{B}$. A *basis matrix* for a subspace $\mathcal{S}$ is any matrix whose columns form a basis for $\mathcal{S}$. A *semi-orthogonal matrix* $\mathbf{A} \in \mathbb{R}^{r\times p}$ has orthogonal columns, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$. A sum of subspaces of $\mathbb{R}^r$ is indicated with the notation '$\oplus$': $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$. For a positive definite matrix $\mathbf{\Sigma} \in \mathbb{S}^{r\times r}$, the inner product in $\mathbb{R}^r$ defined by $\langle \mathbf{x}_1, \mathbf{x}_2\rangle_{\mathbf{\Sigma}} = \mathbf{x}_1^T\mathbf{\Sigma}\mathbf{x}_2$ will be referred to as the $\mathbf{\Sigma}$ inner product; when $\mathbf{\Sigma} = \mathbf{I}_r$, the r by r identity matrix, this inner product will be called the usual inner product. A projection relative to the $\mathbf{\Sigma}$ inner product is the projection operator in the

inner product space $\{\mathbb{R}^r, \langle\cdot,\cdot\rangle_{\mathbf{\Sigma}}\}$; that is, if $\mathbf{B} \in \mathbb{R}^{r\times p}$, then the projection onto span($\mathbf{B}$) relative to $\mathbf{\Sigma}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})} \equiv \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^\dagger\mathbf{B}^T\mathbf{\Sigma}$, where † indicates the Moore-Penrose inverse. The projection onto the orthogonal complement of span($\mathbf{B}$) relative to the $\mathbf{\Sigma}$ inner product, $\mathbf{I}_r - \mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})}$, will be denoted by $\mathbf{Q}_{\mathbf{B}(\mathbf{\Sigma})}$. Projection operators employing the usual inner product will be written with a single subscript argument $\mathbf{P}_{(\cdot)}$, where the subscript describes the subspace, and $\mathbf{Q}_{(\cdot)} = \mathbf{I}_r - \mathbf{P}_{(\cdot)}$. The orthogonal complement $\mathcal{S}^\perp$ of a subspace $\mathcal{S}$ is constructed with respect to the usual inner product, unless indicated otherwise.

2 Envelopes

This article revolves around the parameterization of a covariance matrix in reference to a subspace that contains a conditional mean vector. Specifically, as we saw in (2), this is achieved by decomposing the covariance matrix into the sum of two matrices, each of whose column spaces either contains or is orthogonal to the subspace containing the mean. The only way to do so is to create a split based on the eigenvectors of the covariance. This leads us naturally to invariant and reducing subspaces of a matrix, from which the concept of an envelope arises.

2.1 Invariant and reducing subspaces

Recall that a subspace $\mathcal{R}$ of $\mathbb{R}^r$ is an *invariant subspace* of $\mathbf{M} \in \mathbb{R}^{r\times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$; so $\mathbf{M}$ maps $\mathcal{R}$ to a subset of itself. $\mathcal{R}$ is a *reducing subspace* of $\mathbf{M}$ if, in addition, $\mathbf{M}\mathcal{R}^\perp \subseteq \mathcal{R}^\perp$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$. Some intuition may be provided here by describing how invariant subspaces arise in Zyskind's (1967) pioneering work on linear models. Consider n observations on a univariate linear model written in terms of the $n \times 1$ response vector $\mathbf{W} = \mathbf{F}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\mathbf{F} \in \mathbb{R}^{n\times p}$ is known, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the vector

we would like to estimate and $\mathbf{V} = \text{var}(\boldsymbol{\epsilon}) \in \mathbb{R}^{n\times n}$ denotes the error covariance matrix. The rank of $\mathbf{F}$ may be less than p and $\mathbf{V}$ may be singular. Let $\mathbf{a}^T\boldsymbol{\alpha}$ be an estimable linear combination of the coefficients $\boldsymbol{\alpha}$. Zyskind (1967) showed that the ordinary least squares estimator of $\mathbf{a}^T\boldsymbol{\alpha}$ is equal to the corresponding generalized least squares estimator for every $\mathbf{a} \in \mathbb{R}^p$ if and only if span($\mathbf{F}$) is an invariant subspace of $\mathbf{V}$. Our approach is distinct from Zyskind's since we are working with multivariate models and have quite different goals. Additionally, Zyskind's dimensions grow with n, while ours will remain fixed.

Back to our developments, the next proposition characterizes a matrix $\mathbf{M}$ in terms of projections on its reducing subspaces, and gives exactly the kind of decomposition we are seeking.

Proposition 2.1 *$\mathcal{R}$ reduces $\mathbf{M} \in \mathbb{R}^{r\times r}$ if and only if $\mathbf{M}$ can be written in the form*

$$\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}. \tag{4}$$

Corollary 2.1 describes consequences of Proposition 2.1 (and Lemma A.1 reported in the Supplement), including a relationship between reducing subspaces of $\mathbf{M}$ and $\mathbf{M}^{-1}$, when $\mathbf{M}$ is non-singular.

Corollary 2.1 *Let $\mathcal{R}$ reduce $\mathbf{M} \in \mathbb{R}^{r\times r}$, let $\mathbf{A} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{R}$, and let $\mathbf{A}_0$ be a semi-orthogonal basis matrix for $\mathcal{R}^\perp$. Then*

1. *$\mathbf{M}$ and $\mathbf{P}_{\mathcal{R}}$, and $\mathbf{M}$ and $\mathbf{Q}_{\mathcal{R}}$ commute.*

2. *$\mathcal{R} \subseteq$ span($\mathbf{M}$) if and only if $\mathbf{A}^T\mathbf{M}\mathbf{A}$ is full rank.*

3. *If $\mathbf{M}$ is full rank, then*

$$\mathbf{M}^{-1} = \mathbf{A}(\mathbf{A}^T\mathbf{M}\mathbf{A})^{-1}\mathbf{A}^T + \mathbf{A}_0(\mathbf{A}_0^T\mathbf{M}\mathbf{A}_0)^{-1}\mathbf{A}_0^T. \tag{5}$$

As mentioned in the preamble to this section, there is a connection between the eigenstructure of a symmetric matrix $\mathbf{M}$ and its reducing subspaces. By definition, any invariant subspace of $\mathbf{M} \in \mathbb{S}^{r\times r}$ is also a reducing subspace of $\mathbf{M}$. In particular, it follows from Proposition 2.1 that the subspace spanned by any set of eigenvectors of $\mathbf{M}$ is a reducing subspace of $\mathbf{M}$. This connection is formalized in the following proposition.

Proposition 2.2 *Let $\mathcal{R}$ be a subspace of $\mathbb{R}^r$ and let $\mathbf{M} \in \mathbb{S}^{r\times r}$. Assume that $\mathbf{M}$ has $q \leq r$ distinct eigenvalues, and let $\mathbf{P}_i$, $i = 1, \ldots, q$ indicate the projections on the corresponding eigenspaces. Then the following statements are equivalent:*

1. *$\mathcal{R}$ reduces $\mathbf{M}$,*
2. *$\mathcal{R} = \oplus_{i=1}^{q} \mathbf{P}_i\mathcal{R}$,*
3. *$\mathbf{P}_\mathcal{R} = \sum_{i=1}^{q} \mathbf{P}_i\mathbf{P}_\mathcal{R}\mathbf{P}_i$,*
4. *$\mathbf{M}$ and $\mathbf{P}_\mathcal{R}$ commute.*

2.2 M-envelopes

Since the intersection of two reducing subspaces of a matrix $\mathbf{M} \in \mathbb{S}^{r\times r}$ is itself a reducing subspace, it makes sense to talk about the smallest reducing subspace of $\mathbf{M}$ that contains a certain subspace $\mathcal{S}$, a notion that is central to this article.

Definition 2.1 *Let $\mathbf{M} \in \mathbb{S}^{r\times r}$ and let $\mathcal{S} \subseteq \text{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{S}$, to be written as $\mathcal{E}_\mathbf{M}(\mathcal{S})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{S}$.*

This definition requires that $\mathcal{S} \subseteq \text{span}(\mathbf{M})$. Since the column space of $\mathbf{M}$ is itself a reducing subspace of $\mathbf{M}$, this containment guarantees existence of the $\mathbf{M}$-envelope, and will always be assumed in this article. Note that the containment holds trivially if $\mathbf{M}$ is full rank, i.e. if $\text{span}(\mathbf{M}) = \mathbb{R}^r$. Moreover, closure under intersection guarantees that

the $\mathbf{M}$-envelope is in fact a reducing subspace of $\mathbf{M}$. Thus the $\mathbf{M}$-envelope of $\mathcal{S}$ can be interpreted as the unique smallest reducing subspace of $\mathbf{M}$ that contains $\mathcal{S}$, and represent a well-defined parameter in some statistical problems.

To develop some intuition on $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, consider the case where all the r eigenvalues of $\mathbf{M}$ are distinct. Then, among the 2^r ways of dividing the eigenvectors of $\mathbf{M}$ into two groups, there is one and only one way in which one of the two groups spans a subspace of minimal dimension that contains $\mathcal{S}$. This minimal subspace is $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Thus, in this case, $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is the smallest subspace that contains $\mathcal{S}$ and that is aligned with the eigenstructure of $\mathbf{M}$. Of course, the situation becomes more complicated if $\mathbf{M}$ has less than r distinct eigenvalues, and that is why we use reducing subspaces in the general definition of $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$.

The $\mathbf{M}$-envelope of any reducing subspace is the reducing subspace itself; that is, $\mathcal{E}_{\mathbf{M}}(\mathcal{R}) = \mathcal{R}$ if $\mathcal{R}$ reduces $\mathbf{M}$. A special case of this statement is that, for any subspace $\mathcal{S}$ of $\text{span}(\mathbf{M})$, $\mathcal{E}_{\mathbf{M}}(\mathcal{E}_{\mathbf{M}}(\mathcal{S})) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Thus, as an operator, $\mathcal{E}_{\mathbf{M}}(\cdot)$ is idempotent. Additionally, since an envelope is a reducing subspace, the results in Section 2.1 are applicable.

The following proposition, derived from Proposition 2.2 and Definition 2.1, gives a characterization of $\mathbf{M}$-envelopes.

Proposition 2.3 *Let* $\mathbf{M} \in \mathbb{S}^{r\times r}$*, let* $\mathbf{P}_i$*,* $i = 1, \ldots, q$*, be the projections onto the eigenspaces of* $\mathbf{M}$*, and let* $\mathcal{S}$ *be a subspace of* $\text{span}(\mathbf{M})$*. Then* $\mathcal{E}_{\mathbf{M}}(\mathcal{S}) = \oplus_{i=1}^{q} \mathbf{P}_i\mathcal{S}$*.*

We next investigate how the $\mathbf{M}$-envelope is modified by linear transformations of $\mathcal{S}$. While an envelope does not transform equivariantly for all linear transformations, it does so for symmetric linear transformations that commute with $\mathbf{M}$, as the next proposition shows.

Proposition 2.4 *Let* $\mathbf{K} \in \mathbb{S}^{r\times r}$ *commute with* $\mathbf{M} \in \mathbb{S}^{r\times r}$*, and let* $\mathcal{S}$ *be a subspace of* $\text{span}(\mathbf{M})$*. Then* $\mathbf{K}\mathcal{S} \subseteq \text{span}(\mathbf{M})$ *and the following equivariance holds*

$$\mathcal{E}_{\mathbf{M}}(\mathbf{K}\mathcal{S}) = \mathbf{K}\mathcal{E}_{\mathbf{M}}(\mathcal{S}). \tag{6}$$

If, in addition, $\mathcal{S} \subseteq \text{span}(\mathbf{K})$ *and* $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ *reduces* $\mathbf{K}$*, then the following invariance holds*

$$\mathcal{E}_{\mathbf{M}}(\mathbf{K}\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S}). \tag{7}$$

We conclude this section by exploring a useful consequence of (7). Starting with any function $f : \mathbb{R} \to \mathbb{R}$, we can create $f^* : \mathbb{S}^{r\times r} \to \mathbb{S}^{r\times r}$ as follows. Let m_i and $\mathbf{P}_i$, $i = 1, \ldots, q$ indicate the distinct eigenvalues and the projections on the corresponding eigenspaces for a matrix $\mathbf{M} \in \mathbb{S}^{r\times r}$, and define $f^*(\mathbf{M}) = \sum_{i=1}^{q} f(m_i)\mathbf{P}_i$. If $f(\cdot)$ is such that $f(0) = 0$ and $f(x) \neq 0$ whenever $x \neq 0$, then it is easy to verify that (i) $f^*(\mathbf{M})$ commutes with $\mathbf{M}$, (ii) any subspace $\mathcal{S} \subseteq \text{span}(\mathbf{M})$ is also $\mathcal{S} \subseteq \text{span}\{f^*(\mathbf{M})\}$, and (iii) $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ reduces $f^*(\mathbf{M})$. Hence, by Proposition 2.4 we have $\mathcal{E}_{\mathbf{M}}(f^*(\mathbf{M})\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$. In particular, this guarantees invariance for any power of $\mathbf{M}$:

$$\mathcal{E}_{\mathbf{M}}(\mathbf{M}^k\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S}) \text{ for all } k \in \mathbb{R}. \tag{8}$$

3 Envelope Models

3.1 Theoretical formulation of envelope models

We are now in a position to refine model (1) by using an envelope to connect $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Let $\mathcal{B} = \text{span}(\boldsymbol{\beta})$, $d = \dim(\mathcal{B})$ and, to exclude the trivial case, assume $d > 0$. Consider the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, of dimension u, so that $0 < d \leq u \leq r$. We use this envelope as a well-defined parameter to link the mean and variance structures of the multivariate linear model. Since $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is unknown, it needs to be estimated, and this is facilitated by writing formal model statements that incorporate it as a parameter. We give two such statements: A coordinate-free version that uses $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ as the parameter, and a coordinate version that uses a semi-orthogonal basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Both versions have

advantages, depending on the phase of the analysis. For instance, the coordinate version is necessary for computation. Our use of "coordinate-free" and "coordinate" terminology applies only to the representation of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, and not to the rest of the model.

Since $\mathbf{\Sigma}$ is a positive definite matrix reduced by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, all of the results in Section 2 apply. In particular, $\mathbf{\Sigma}$ can be written in the form given by Proposition 2.1 with $\mathcal{R} = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, its inverse can be expressed as in part 3 of Corollary 2.1, and $\mathbf{\Sigma}^k\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathbf{\Sigma}^k\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ for all $k \in \mathbb{R}$, because of Proposition 2.4. The following corollary gives a coordinate-free version of Proposition 2.1, making use of the additional properties characterizing a covariance matrix.

Corollary 3.1 *A subspace $\mathcal{R}$ of $\mathbb{R}^r$ reduces $\mathbf{\Sigma}$ if and only if $\mathbf{\Sigma}$ can be written in the form $\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are symmetric positive semi-definite matrices such that $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{0}$ and $\mathcal{R} = \text{span}(\mathbf{\Sigma}_1)$.*

The coordinate-free representation of the envelope model is model (1) with error covariance matrix satisfying

$$\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2,\ \mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{0},\ \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \text{span}(\mathbf{\Sigma}_1). \tag{9}$$

Since reducing subspaces are specified by this decomposition of $\mathbf{\Sigma}$, we could equivalently replace the requirement $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \text{span}(\mathbf{\Sigma}_1)$ with the condition that $\text{span}(\mathbf{\Sigma}_1)$ has minimal dimension under the constraint $\mathcal{B} \subseteq \text{span}(\mathbf{\Sigma}_1)$. However, it is important to note that (9), *per se*, does not restrict the scope of model (1). If $u = r$, then we must have $\mathbf{\Sigma}_1 = \mathbf{\Sigma}$ and $\mathbf{\Sigma}_2 = \mathbf{0}$. If $r \leq p$ and $d = r$, then the envelope model coincides with the standard multivariate linear model, since there are evidently no linear redundancies in (1), and thus no reduction is possible with the new parameterization. On the other hand, if $u < r$ then there is a potential for the envelope model expressed through (9) to yield substantial gains. As an extension of the ideas presented here, alternative uses of envelopes that allow

reduction when $r \leq p$ and $d = r$ are described in Section 8.4.

To write the coordinate version of the envelope model, let $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, and let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix. Then there is an $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ such that $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$. Additionally, let $\mathbf{\Omega} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma} \in \mathbb{S}^{u\times u}$ and let $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0 \in \mathbb{S}^{(r-u)\times(r-u)}$. Then, using Proposition 2.1 and Corollary 3.1 we can write

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\alpha} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \\ \mathbf{\Sigma} &= \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \end{aligned} \tag{10}$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean $\mathbf{0}$ and variance $\mathbf{\Sigma}$. The matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ can be thought of as *coordinate matrices*, since they carry the coordinates of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ relative to $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$, just as $\boldsymbol{\eta}$ contains the coordinates of $\boldsymbol{\beta}$ relative to $\mathbf{\Gamma}$.

The total number N of parameters needed to estimate (10) is

$$N = r + pu + u(r-u) + \frac{u(u+1)}{2} + \frac{(r-u)(r-u+1)}{2}.$$

The first term on the right hand side corresponds to the intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$. The second term corresponds to the unconstrained coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$. The last two terms correspond to $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$. Their parameter counts arise because, for any integer $k > 0$, it takes $k(k+1)/2$ numbers to specify a nonsingular matrix in $\mathbb{S}^{k\times k}$. The third term, $u(r-u)$, which corresponds roughly to $\mathbf{\Gamma}$, arises as follows. The matrix $\mathbf{\Gamma}$ is not identified, since for any orthogonal matrix $\mathbf{A}$ replacing $\mathbf{\Gamma}$ with $\mathbf{\Gamma}\mathbf{A}$ results in an equivalent model. However, $\text{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is identified and estimable. The parameter space for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is a Grassmann manifold $\mathbb{G}^{r\times u}$ of dimension u in $\mathbb{R}^r$; that is, the collection of all u-dimensional subspaces of $\mathbb{R}^r$. From basic properties of Grassmann manifolds it is known that $u(r-u)$ parameters are needed to specify an element of $\mathbb{G}^{r\times u}$ (Edelman, Tomás and

Smith, 1998). Once $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is determined, so is its orthogonal complement span($\mathbf{\Gamma}_0$), and no additional free parameters are required.

Simplifying the above expression for N, we obtain $N = r + pu + r(r+1)/2$. The difference between the total parameter count for the full model (1) with $r = u$ and the envelope model (10) with $u < r$ is therefore $p(r-u)$.

Note that a specific envelope model is identified by the value of u, with the full model (1) occurring when $u = r$. All envelope models are nested within the full model, but two envelope models with different values of u are not necessarily nested. To see this, it is enough to realize that the number of free parameters needed to specify an element of $\mathbb{G}^{r\times u}$ is the same for $u = 1$ and $u = r - 1$. In full generality, u is a model selection parameter that can be chosen using traditional reasoning, as discussed in Section 7.1.

3.2 Alternative envelopes for random designs

The models introduced so far are parameterized in terms of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, the $\mathbf{\Sigma}$-envelope of $\mathcal{B}$, in coordinate-free and coordinate versions. While this seems to be the natural route when $\mathbf{X}$ is chosen by design, other choices are available when $\mathbf{X}$ is random. For instance, we might create a parameterization in terms of $\mathcal{E}_{\mathbf{\Sigma_Y}}(\mathcal{B})$, the envelope of $\mathcal{B}$ based on the marginal response covariance matrix $\mathbf{\Sigma_Y} = \mathrm{var}(\mathbf{Y})$. The next proposition states equality of several envelopes. The first equality shows an important equivalence between enveloping in reference to the error variability $\mathbf{\Sigma}$ and the response variability $\mathbf{\Sigma_Y}$. The other equalities will be relevant in Section 8.

Proposition 3.1 *Assume model (1). Then* $\mathbf{\Sigma}^{-1}\mathcal{B} = \mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}$, *and*

$$\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma_Y}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1}\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma_Y}}(\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma_Y}}(\mathbf{\Sigma}^{-1}\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}).$$

4 Maximum Likelihood Estimation

Before deriving the MLEs for the envelope model, we give a few preliminary results in Section 4.1. These are intended primarily to facilitate derivations in Section 4.2 but, like the results in Section 2, may have wider applicability. The calculations necessary to obtain the estimates are summarized in Section 4.3.

4.1 Preliminary results

Lemma 4.1 *Let* $\mathbf{U} \in \mathbb{R}^{n\times p}$, $\mathbf{V} \in \mathbb{R}^{n\times r}$, *and* $\mathbf{W} \in \mathbb{R}^{p\times d}$ *be known matrices. Let* $\mathbf{\Lambda}$ *be a positive semi-definite matrix in* $\mathbb{R}^{p\times p}$ *such that* $\text{span}(\mathbf{W}) \subseteq \text{span}(\mathbf{\Lambda})$. *Then the minimizer of*

$$\text{tr}\left[(\mathbf{U}-\mathbf{A})\mathbf{\Lambda}(\mathbf{U}-\mathbf{A})^T\right] \tag{11}$$

over the set of matrices $\mathcal{A} = \{\mathbf{A} : \text{ span}(\mathbf{A}) \subseteq \text{span}(\mathbf{V}),\ \text{span}(\mathbf{A}^T) \subseteq \text{span}(\mathbf{W})\}$ *is* $\mathbf{A}^* = \mathbf{P}_\mathbf{V}\mathbf{U}\mathbf{P}^T_{\mathbf{W}(\mathbf{\Lambda})}$, *and the corresponding minimum of (11) is*

$$\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T) - \text{tr}(\mathbf{P}_\mathbf{V}\mathbf{U}\mathbf{P}^T_{\mathbf{W}(\mathbf{\Lambda})}\mathbf{\Lambda}\mathbf{P}_{\mathbf{W}(\mathbf{\Lambda})}\mathbf{U}^T\mathbf{P}_\mathbf{V}).$$

For a nonzero $\mathbf{A} \in \mathbb{S}^{r\times r}$ (i.e. an $r \times r$ symmetric matrix whose entries are not all equal to 0), we denote by $\det_0(\mathbf{A})$ the product of its non-zero eigenvalues. Note that, for any constant c, $\det_0(c\mathbf{A}) = c^k\det_0(\mathbf{A})$, where k is the rank of $\mathbf{A}$. The next lemma will facilitate analysis with the structure introduced in Corollary 3.1.

Lemma 4.2 *If* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are nonzero symmetric matrices such that* $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$, *then*

1. $\det_0(\mathbf{A}_1 + \mathbf{A}_2) = \det_0(\mathbf{A}_1) \times \det_0(\mathbf{A}_2)$,
2. $(\mathbf{A}_1 + \mathbf{A}_2)^\dagger = \mathbf{A}_1^\dagger + \mathbf{A}_2^\dagger$, *and*

3. $(\mathbf{A}_1 + \mathbf{A}_2)^r = \mathbf{A}_1^r + \mathbf{A}_2^r$, *for any* $r > 0$.

Finally, we introduce a lemma that gives an explicit expression for the MLE of the covariance matrix in a multivariate normal likelihood, when the column space of the covariance is fixed and the mean is known.

Lemma 4.3 *Let* $\mathcal{A}$ *be a class of* $p \times p$ *positive semi-definite matrices having the same column space of dimension* k, $0 < k \leq p$, *and* $\mathbf{P}$ *be the projection onto the common column space. Let* $\mathbf{U}$ *be a matrix in* $\mathbb{R}^{n\times p}$, *and let*

$$L(\mathbf{A}) = [\det{}_0(\mathbf{A})]^{-\frac{1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{U}\mathbf{A}^\dagger\mathbf{U}^T)}.$$

Then the maximizer of $L(\mathbf{A})$ *over* $\mathcal{A}$ *is the matrix* $n^{-1}\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P}$, *and the maximum value of* $L(\mathbf{A})$ *is* $n^{k/2}e^{-nk/2}[\det{}_0(\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P})]^{-1/2}$.

4.2 Coordinate-free representation of the MLE

Derivation of the MLE is easier using the coordinate-free representation of the envelope model, as given by (1) and (9). We assume that the observations $\mathbf{Y}_i$, $i = 1, \ldots, n$, are independent, and that $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y}|\mathbf{X}_i$, $i = 1, \ldots, n$, with $\bar{\mathbf{X}} = 0$. We assume also that $n > r + p$. Let $\mathbf{G}$ be the $n \times r$ matrix whose ith row is $\mathbf{Y}_i^T$, $\mathbf{F}$ be the $n \times p$ matrix whose ith row is $\mathbf{X}_i^T$, and $\mathbf{1}_n$ be the $n \times 1$ vector with each entry equal to 1.

For a $\mathbf{\Sigma}$-envelope with fixed dimension u, $0 < u < r$, the likelihood based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is

$$\begin{aligned} L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) &= [\det(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)]^{-\frac{1}{2}} \\ &\times \mathrm{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)^T], \end{aligned} \tag{12}$$

where etr$(\cdot)$ denotes the composite function $\exp \circ \operatorname{tr}(\cdot)$, and $\otimes$ the Kronecker product. This likelihood is to be maximized over $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ subject to the constraints:

$$\operatorname{span}(\boldsymbol{\beta}) \subseteq \operatorname{span}(\boldsymbol{\Sigma}_1), \quad \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2 = \mathbf{0}. \tag{13}$$

By Lemma 4.2, and using the relation $\boldsymbol{\Sigma}_2 \boldsymbol{\beta} = \mathbf{0}$, the likelihood in (12) can be factored as $L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1) \times L_2^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_2)$, where

$$\begin{aligned} L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1) = \; & [\det{}_0(\boldsymbol{\Sigma}_1)]^{-1/2} \\ & \times \operatorname{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)^T] \\ L_2^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_2) = \; & [\det{}_0(\boldsymbol{\Sigma}_2)]^{-1/2} \times \operatorname{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n)\boldsymbol{\Sigma}_2^{\dagger}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n)^T]. \end{aligned} \tag{14}$$

Based on this factorization and the constraints in (13), we can decompose the likelihood maximization into the following steps:

1. Fix $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\beta}$, and maximize $L^{(u)}$ in (12) over $\boldsymbol{\alpha}$; then substitute the optimal $\boldsymbol{\alpha}$ into $L_1^{(u)}$ and $L_2^{(u)}$ in (14) to obtain $L_{11}^{(u)}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1)$ and $L_{21}^{(u)}(\boldsymbol{\Sigma}_2)$. The required maximizer is the sample mean of $\{\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i : i = 1, \ldots, n\}$ which, because $\mathbf{X}$ has sample mean zero, is simply $\bar{\mathbf{Y}}$. Hence, if we let $\mathbf{U}$ be the $n \times r$ matrix whose ith row is $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$, the partially maximized $L_1^{(u)}$ and $L_2^{(u)}$ are

$$\begin{aligned} L_{11}^{(u)}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1) = & [\det{}_0(\boldsymbol{\Sigma}_1)]^{-1/2} \times \operatorname{etr}[-\frac{1}{2}(\mathbf{U} - \mathbf{F}\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{U} - \mathbf{F}\boldsymbol{\beta}^T)^T], \\ L_{21}^{(u)}(\boldsymbol{\Sigma}_2) = & [\det{}_0(\boldsymbol{\Sigma}_2)]^{-1/2} \times \operatorname{etr}(-\frac{1}{2}\mathbf{U}\boldsymbol{\Sigma}_2^{\dagger}\mathbf{U}^T). \end{aligned} \tag{15}$$

2. Fix $\boldsymbol{\Sigma}_1$, and further maximize the function $L_{11}^{(u)}$ from step 1 over $\boldsymbol{\beta}$, subject to the first constraint in (13), to obtain $L_{12}^{(u)}(\boldsymbol{\Sigma}_1)$. For this maximization we use

Lemma 4.1, with the relevant quadratic form given by

$$\mathrm{tr}[(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_1^\dagger(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T)^T] \equiv \mathrm{tr}[(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r)\boldsymbol{\Sigma}_1^\dagger(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r)^T].$$

Thus, the optimal $\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r$ is $\mathbf{P}_{\mathbf{F}}\mathbf{U}\mathbf{P}^T_{\mathbf{I}_r(\boldsymbol{\Sigma}_1^\dagger)} = \mathbf{P}_{\mathbf{F}}\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_1}$. This implies that

$$\boldsymbol{\beta} = \mathbf{P}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}, \tag{16}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}} = \mathbf{U}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$ is the MLE of $\boldsymbol{\beta}$ from the full model (1). Consequently, we see that $\widehat{\boldsymbol{\beta}}$ will be the projection of $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$ onto the MLE of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Substituting this into (15), and using the relation $\mathbf{P}_{\boldsymbol{\Sigma}_1}\boldsymbol{\Sigma}_1^\dagger = \boldsymbol{\Sigma}_1^\dagger$, we see that the maximum of $L_{11}^{(u)}(\boldsymbol{\beta},\boldsymbol{\Sigma}_1)$ for fixed $\boldsymbol{\Sigma}_1$ over $\boldsymbol{\beta}$ is

$$\begin{aligned} L_{12}^{(u)}(\boldsymbol{\Sigma}_1) =& [\det{}_0(\boldsymbol{\Sigma}_1)]^{-1/2} \times \mathrm{etr}[-\frac{1}{2}(\mathbf{U}-\mathbf{P}_{\mathbf{F}}\mathbf{U})\boldsymbol{\Sigma}_1^\dagger(\mathbf{U}-\mathbf{P}_{\mathbf{F}}\mathbf{U})^T] \\ =& [\det{}_0(\boldsymbol{\Sigma}_1)]^{-1/2} \times \mathrm{etr}(-\frac{1}{2}\mathbf{Q}_{\mathbf{F}}\mathbf{U}\boldsymbol{\Sigma}_1^\dagger\mathbf{U}^T\mathbf{Q}_{\mathbf{F}}), \end{aligned} \tag{17}$$

where $\mathbf{Q}_{\mathbf{F}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{F}}$.

3. Using Lemma 4.3, maximize $L_{12}^{(u)}(\boldsymbol{\Sigma}_1)$ over all $\boldsymbol{\Sigma}_1$'s having the same column space, to obtain $L_{13}^{(u)}(\mathbf{P}_{\boldsymbol{\Sigma}_1})$, which is proportional to $\left[\det{}_0(\mathbf{P}_{\boldsymbol{\Sigma}_1}\mathbf{U}^T\mathbf{Q}_{\mathbf{F}}\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_1})\right]^{-1/2}$. Similarly, maximize $L_{21}^{(u)}(\boldsymbol{\Sigma}_2)$ over all $\boldsymbol{\Sigma}_2$'s having the same column space, to obtain $L_{22}^{(u)}(\mathbf{P}_{\boldsymbol{\Sigma}_2})$, which is proportional to $\left[\det{}_0(\mathbf{P}_{\boldsymbol{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_2})\right]^{-1/2}$. Note that $L_{13}^{(u)}$ depends only on the column space of $\boldsymbol{\Sigma}_1$, and $L_{22}^{(u)}$ only on the column space of $\boldsymbol{\Sigma}_2$.

4. Optimize the partially maximized likelihood $L_{13}^{(u)}(\mathbf{P}_{\boldsymbol{\Sigma}_1}) \times L_{22}^{(u)}(\mathbf{P}_{\boldsymbol{\Sigma}_2})$, which is pro-

portional to

$$[\det_0(\mathbf{P}_{\boldsymbol{\Sigma}_1}\mathbf{U}^T\mathbf{Q}_\mathbf{F}\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_1})]^{-1/2} \times [\det_0(\mathbf{P}_{\boldsymbol{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_2})]^{-1/2}$$
$$= [\det_0(\mathbf{P}_{\boldsymbol{\Sigma}_1}\mathbf{U}^T\mathbf{Q}_\mathbf{F}\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_1} + \mathbf{P}_{\boldsymbol{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\boldsymbol{\Sigma}_2})]^{-1/2}. \quad (18)$$

Because $\mathbf{P}_{\boldsymbol{\Sigma}_2} = \mathbf{I}_r - \mathbf{P}_{\boldsymbol{\Sigma}_1} = \mathbf{Q}_{\boldsymbol{\Sigma}_1}$, the above depends on $\mathbf{P}_{\boldsymbol{\Sigma}_1}$ alone. Additionally, $\mathbf{U}^T\mathbf{U}$ is n times the marginal sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_\mathbf{Y}$ of the responses, and $\mathbf{U}^T\mathbf{Q}_\mathbf{F}\mathbf{U}$ is n times the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{res}}$ of the residuals from the fit of the full model (1). Since we have assumed that $n > r + p$, it follows that $\text{rank}(\widehat{\boldsymbol{\Sigma}}_{\text{res}}) = \text{rank}(\widehat{\boldsymbol{\Sigma}}_\mathbf{Y}) = r$ with probability 1. Therefore $\det_0(\cdot)$ in (18) can be replaced by $\det(\cdot)$, the usual determinant, and we need to minimize the function

$$D = D(\text{span}(\boldsymbol{\Sigma}_1)) \equiv \det(\mathbf{P}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\Sigma}}_{\text{res}}\mathbf{P}_{\boldsymbol{\Sigma}_1} + \mathbf{Q}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\Sigma}}_\mathbf{Y}\mathbf{Q}_{\boldsymbol{\Sigma}_1}). \quad (19)$$

over the Grassmann manifold $\mathbb{G}^{r\times u}$, subject to the constraint that $\text{rank}(\mathbf{P}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\Sigma}}_{\text{res}}\mathbf{P}_{\boldsymbol{\Sigma}_1}) = u$ – which arises because $\text{rank}(\boldsymbol{\Sigma}_1) = u < r$.

4.3 Implementation of the MLE

The MLE described in Section 4.2 hinges on being able to minimize $\log D$ over the Grassmann manifold $\mathbb{G}^{r\times u}$, where D is as defined in (19). Available gradient-based algorithms for Grassmann optimization (see Edelman, Tomás and Smith, 1998; Liu, Srivastava and Gallivan, 2004) require a coordinate version of the objective function which must have continuous directional derivatives. A coordinate version of objective function (19) satisfies this continuity requirement when $\boldsymbol{\Sigma} > \mathbf{0}$. Recall that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ are semi-orthogonal basis matrices of $\text{span}(\boldsymbol{\Sigma}_1) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and its orthogonal complement, respectively. Let $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Gamma}}_0$ be semi-orthogonal bases for $\text{span}(\widehat{\boldsymbol{\Sigma}}_1)$ and its orthogonal complement. Then $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\beta}}_{\text{fm}}$, $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Sigma}}_\mathbf{Y}\widehat{\boldsymbol{\Gamma}}_0$. Since $\widehat{\boldsymbol{\Sigma}}_{\text{res}}$ and $\widehat{\boldsymbol{\Sigma}}_\mathbf{Y}$

have rank r almost surely, the matrices $\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\boldsymbol{\Gamma}_0$ are positive definite almost surely. Let $\log\det(\cdot)$ denote the composite function $\log\circ\det(\cdot)$. Then the coordinate form of $\log D$ is

$$\begin{aligned}\log D &= \log\det[\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + (\mathbf{I}_r - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}(\mathbf{I}_r - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)] \\ &= \log\det(\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\boldsymbol{\Gamma}) + \log\det(\boldsymbol{\Gamma}_0^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\boldsymbol{\Gamma}_0). \qquad (20)\end{aligned}$$

In summary, maximum likelihood estimation for the parameters involved in the envelope model can be implemented as follows:

a. Obtain the sample version $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$ of the marginal covariance matrix of $\mathbf{Y}$, and obtain the residual covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{res}}$ and the MLE $\widehat{\boldsymbol{\beta}}_{\text{fm}}$ of $\boldsymbol{\beta}$ from the fit of the full model (1).

b. Estimate $\mathbf{P}_{\boldsymbol{\Sigma}_1}$ by minimizing the objective function (20) over the Grassmann manifold $\mathbb{G}^{r\times u}$, and denote the result by $\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}$. Estimate $\mathbf{P}_{\boldsymbol{\Sigma}_2}$ by $\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_2} = \mathbf{I}_r - \widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}$.

c. Estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\beta}}_{\text{fm}}$.

d. Estimate $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ by $\widehat{\boldsymbol{\Sigma}}_1 = \widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}$ and $\widehat{\boldsymbol{\Sigma}}_2 = (\mathbf{I}_r - \widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1})\widehat{\boldsymbol{\Sigma}}_Y(\mathbf{I}_r - \widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1})$.

We assumed at the outset of this derivation that $u < r$. If $u = r$ then $\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1} = \mathbf{I}_r$ and $\widehat{\boldsymbol{\beta}}$ reduces to the usual MLE based on (1). Generally, objective functions defined on Grassmann manifolds can have multiple local optima, but we have not noticed local minima to be an issue for (20).

5 Asymptotic Variances

There is a multitude of approaches for dealing with dimensionality issues in multivariate regression. Many of these, ranging from various versions of principal components to

a multivariate implementation of sliced inverse regression (Li, Aragon, Shedden and Agnan, 2003) are algorithmic in nature, making it difficult to determine post-application standard errors and other inference-related quantities. Unlike these approaches, our analysis of envelope models is based entirely on the likelihood. We are therefore able to pursue inference classically, with methodology that inherits optimal properties from general likelihood theory.

5.1 Estimable functions

The parameters in the coordinate representation (10) of the envelope model can be combined into the vector

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\eta}) \\ \mathrm{vec}(\boldsymbol{\Gamma}) \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \end{pmatrix} \tag{21}$$

where the "vector" operator vec : $\mathbb{R}^{r\times p} \to \mathbb{R}^{rp}$ stacks the columns of the argument matrix. On the symmetric matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ we use the related "vector half" operator vech : $\mathbb{S}^{r\times r} \to \mathbb{R}^{r(r+1)/2}$, which extracts their unique elements (vech stacks only the unique part of each column that lies on or below the diagonal). vec and vech are related through a "contraction" matrix $\mathbf{C}_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ and an "expansion" matrix $\mathbf{E}_r \in \mathbb{R}^{r^2\times r(r+1)/2}$, which are defined so that $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_r\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_r\mathrm{vech}(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{S}^{r\times r}$. These relations uniquely define $\mathbf{C}_r$ and $\mathbf{E}_r$, and imply $\mathbf{C}_r\mathbf{E}_r = \mathbf{I}_{r(r+1)/2}$. For further background on these operators, see Henderson and Searle (1979).

Selected elements of $\boldsymbol{\phi}$ might be of interest in some applications, but here we focus

on some specific estimable functions under the envelope model:

$$\mathbf{h}(\boldsymbol{\phi}) \equiv \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \text{vec}(\boldsymbol{\Gamma}\boldsymbol{\eta}) \\ \text{vech}(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) \end{pmatrix} \equiv \begin{pmatrix} \mathbf{h}_1(\boldsymbol{\phi}) \\ \mathbf{h}_2(\boldsymbol{\phi}) \end{pmatrix}.$$

We have neglected the intercept $\boldsymbol{\alpha}$ in this setup. This induces no loss of generality because the intercept is not involved in $\mathbf{h}$, and its maximum likelihood estimate is asymptotically independent of the other parameter estimates.

If the gradient matrix

$$\mathbf{H} = \begin{pmatrix} \partial\mathbf{h}_1/\partial\boldsymbol{\phi}_1^T & \cdots & \partial\mathbf{h}_1/\partial\boldsymbol{\phi}_4^T \\ \partial\mathbf{h}_2/\partial\boldsymbol{\phi}_1^T & \cdots & \partial\mathbf{h}_2/\partial\boldsymbol{\phi}_4^T \end{pmatrix} \tag{22}$$

were of full rank when evaluated at the true parameter values, then standard methods could be used to find the asymptotic covariance matrices for $\hat{\mathbf{h}}_1 = \mathbf{h}_1(\hat{\boldsymbol{\phi}})$ and $\hat{\mathbf{h}}_2 = \mathbf{h}_2(\hat{\boldsymbol{\phi}})$. However, because of the over-parameterization in $\boldsymbol{\Gamma}$, $\mathbf{H}$ is not of full rank, and standard methods do not apply directly. Nevertheless, $\mathbf{h}$ is identified and estimable, which enables us to use a result by Shapiro (1986, Proposition 4.1) to derive the asymptotic distribution and efficiency gain of the envelope model, as given by the following theorem.

Theorem 5.1 *Suppose* $\bar{\mathbf{X}} = \mathbf{0}$. *Let* $\mathbf{J}$ *be the Fisher information for* $(\text{vec}^T(\boldsymbol{\beta}), \text{vech}^T(\boldsymbol{\Sigma}))^T$ *in the full model (1):*

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n$, *and let* $\mathbf{V} = \mathbf{J}^{-1}$ *be the asymptotic variance of the*

MLE under the full model. Then

$$\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{V}_0), \tag{23}$$

where $\mathbf{V}_0 = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$ *and* $\mathbf{H}$ *is given by*

$$\begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{(r-u)} \end{pmatrix}. \tag{24}$$

Moreover, $\mathbf{V}^{-\frac{1}{2}}(\mathbf{V} - \mathbf{V}_0)\mathbf{V}^{-\frac{1}{2}} = \mathbf{Q}_{\mathbf{J}^{\frac{1}{2}}\mathbf{H}} \geq 0$*, so the envelope model decreases the asymptotic variance by the fraction* $\mathbf{Q}_{\mathbf{J}^{\frac{1}{2}}\mathbf{H}}$.

We next present an alternative form for $\mathbf{V}_0$ that may facilitate computing and that will be helpful in the next section. Since $\mathbf{V}_0$ in (23) depends only on the column space of $\mathbf{H}$ we can replace $\mathbf{H}$ by any matrix $\mathbf{H}_1$ that has the same column space as $\mathbf{H}$. The most convenient and interpretable choice of $\mathbf{H}_1$ is one that makes $\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1$ block-diagonal, with blocks corresponding to the parameters in (21). We now give such a construction. Let $\mathbf{H}_1$ be the $\{pr + r(r+1)/2\} \times \{pu + r(r+1)/2\}$ matrix

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} \end{pmatrix} \\ &\equiv \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} \end{pmatrix}, \end{aligned} \tag{25}$$

and let $\mathbf{H}_2$ be the $\{pu + r(r+1)/2\} \times \{pu + r(r+1)/2 + u^2\}$ matrix whose blocks conform

to those of $\mathbf{H}_1$:

$$\mathbf{H}_2 = \begin{pmatrix} \mathbf{I}_{pu} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_u \otimes \boldsymbol{\Gamma}_0^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_u(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}^T) & \mathbf{I}_{u(u+1)/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{(r-u)(r-u+1)/2} \end{pmatrix}.$$

Then, by direct computation using the relation $\mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}) = \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u\mathbf{C}_u$ (Henderson and Searle, 1979), we have $\mathbf{H} = \mathbf{H}_1\mathbf{H}_2$. Because $\mathbf{H}_2$ has full row rank, we have $\text{span}(\mathbf{H}) = \text{span}(\mathbf{H}_1)$. Furthermore, by straightforward multiplication we see that $\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1$ is the desired block-diagonal matrix. Thus, we can now write

$$\mathbf{V}_0 = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1)^\dagger\mathbf{H}_1^T = \sum_{j=1}^{4} \mathbf{H}_{1j}(\mathbf{H}_{1j}^T\mathbf{J}\mathbf{H}_{1j})^\dagger\mathbf{H}_{1j}^T. \tag{26}$$

5.2 Regression coefficients

Henceforth we write an asymptotic covariance matrix as $\text{avar}(\cdot)$; that is, if $\sqrt{n}(\mathbf{T} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{A})$, then $\text{avar}(\sqrt{n}\,\mathbf{T}) = \mathbf{A}$. We now focus our attention on the asymptotic covariance matrix $\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})]$ of the estimate $\text{vec}(\hat{\boldsymbol{\beta}})$ of $\text{vec}(\boldsymbol{\beta})$ under the envelope model, since this will likely be of most use in practice. This matrix is the upper $pr \times pr$ block diagonal of $\mathbf{V}_0 = \text{avar}(\sqrt{n}\hat{\mathbf{h}})$. Since the first blocks of $\mathbf{H}_{13}$ and $\mathbf{H}_{14}$ are both $\mathbf{0}$, we have (see Supplement, Section D)

$$\begin{aligned} \text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})] &= (\mathbf{I}_p \otimes \boldsymbol{\Gamma})(\mathbf{H}_{11}^T\mathbf{J}\mathbf{H}_{11})^\dagger(\mathbf{I}_p \otimes \boldsymbol{\Gamma}^T) + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)(\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12})^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T) \\ &= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)(\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12})^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T), \end{aligned} \tag{27}$$

where $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12} = \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_u \otimes \mathbf{I}_{r-u}$. If $u = r$, then $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$ and the second term on the right hand side of (27) does not appear. The first

term on the right hand side of (27) is the asymptotic variance for $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\Gamma}$ is known, and the second term can be interpreted as the "cost" of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The total on the right does not exceed $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$, which is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ from the full model. A transparent decomposition of this asymptotic variance will be given in the next section.

Although we do not have a full proof, we expect that $\mathbf{H}_{12}^T \mathbf{J} \mathbf{H}_{12}$ will be of full rank, so that regular inverses can be used. This expectation is based on the following reasoning for two extreme cases. Suppose that $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ have no eigenvalues in common. Then it can be shown that $(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_u \otimes \mathbf{I}_{r-u}) > \mathbf{0}$. Since $\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} \geq \mathbf{0}$, it follows that $\mathbf{H}_{12}^T \mathbf{J} \mathbf{H}_{12} > \mathbf{0}$. On the other extreme, suppose that $\boldsymbol{\Omega} = \mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = \mathbf{I}_{r-u}$, so that all their eigenvalues are identical. Then $\mathbf{H}_{12}^T \mathbf{J} \mathbf{H}_{12} = \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \mathbf{I}_{r-u}$, but in this case $\boldsymbol{\eta}$ must have full row rank equal to d, and again $\mathbf{H}_{12}^T \mathbf{J} \mathbf{H}_{12} > \mathbf{0}$.

5.3 Fitted values and predictions

From the above asymptotic results we can derive the asymptotic distribution of the fitted values, as well as the asymptotic prediction variance. In our context the fitted values at a particular $\mathbf{X}$ can be written as $\hat{\mathbf{Y}} = \hat{\boldsymbol{\beta}}\mathbf{X} = (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{vec}(\hat{\boldsymbol{\beta}})$. Hence the fitted value $\hat{\mathbf{Y}}$ has the following asymptotic distribution

$$\sqrt{n}(\hat{\mathbf{Y}} - \mathrm{E}(\hat{\mathbf{Y}})) \xrightarrow{\mathcal{L}} N(0, (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})](\mathbf{X} \otimes \mathbf{I}_r)). \tag{28}$$

The asymptotic mean squared error for prediction at $\mathbf{X}$ can be deduced similarly. Suppose that, at some value of $\mathbf{X}$, we observe a new $\mathbf{Y}$ – independently of the past observations

$(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$. Then

$$\begin{aligned}
&\mathrm{E}[(\hat{\mathbf{Y}} - \mathbf{Y})(\hat{\mathbf{Y}} - \mathbf{Y})^T] \\
&\quad = \mathrm{E}[(\hat{\mathbf{Y}} - \mathrm{E}(\hat{\mathbf{Y}}))(\hat{\mathbf{Y}} - \mathrm{E}(\hat{\mathbf{Y}}))^T] + \mathrm{E}[(\mathrm{E}(\hat{\mathbf{Y}}) - \mathbf{Y})(\mathrm{E}(\hat{\mathbf{Y}}) - \mathbf{Y})^T],
\end{aligned}$$

where the cross-product terms vanish because $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are independent. Combining this with expression (28), we see that the mean squared error of the prediction is approximated by

$$\mathrm{E}[(\hat{\mathbf{Y}} - \mathbf{Y})(\hat{\mathbf{Y}} - \mathbf{Y})^T] = n^{-1}(\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})](\mathbf{X} \otimes \mathbf{I}_r) + \boldsymbol{\Sigma} + o(n^{-1}).$$

6 Interpretations

To gain further insight into the structure of our envelope model for multivariate linear regression, we now provide interpretations for the various quantities in the asymptotic variance of $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})$ derived in the last section. The key to understanding this variance structure is the special structure of the joint Fisher information for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_4^T)^T$, as defined in (21). Let $\ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)$ denote the likelihood function for the $\boldsymbol{\phi}$'s. We will adopt the following notation:

$$\mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}} = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_1^T}\right], \quad \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\Gamma}} = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_2^T}\right], \tag{29}$$

and so on. Although it may be more technically correct to use notation such as $\mathbf{J}_{\boldsymbol{\phi}_1\boldsymbol{\phi}_2}$, we will nevertheless use (29) to keep track of the original parameters. Furthermore, we will use notations such as $\mathbf{J}_{(\boldsymbol{\eta},\boldsymbol{\Gamma})(\boldsymbol{\eta},\boldsymbol{\Gamma})}$ to denote the joint information for parameter sub-vectors such as $(\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)^T$.

From the discussion in Section 5 it can be deduced that the Fisher information for

$(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)$, $\mathbf{H}^T\mathbf{J}\mathbf{H}$, has the following form:

$$\begin{pmatrix} \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\Gamma}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0} \end{pmatrix}, \tag{30}$$

with specific expressions for the non-zero blocks given in the Supplement, Section E. What is special about this form is that, if we cross out the second row and second column, the remaining matrix is block-diagonal with three diagonal blocks; $\mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}}$, $\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Omega}}$, and $\mathbf{J}_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0}$. Similarly, if we cross out the first row and first column, the remaining matrix is block-diagonal with two diagonal blocks; $\mathbf{J}_{(\boldsymbol{\Gamma},\boldsymbol{\Omega})(\boldsymbol{\Gamma},\boldsymbol{\Omega})}$ and $\mathbf{J}_{\boldsymbol{\Omega}_0}$. This implies two important facts:

1. If $\boldsymbol{\Gamma}$ is known, then the asymptotic variance of the MLE of $\boldsymbol{\eta}$, say $\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$, is simply $\mathbf{J}^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}$. The other two parameters, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, have no plugging-in effect.

2. If $\boldsymbol{\eta}$ is known, then the asymptotic variance of the MLE of $\boldsymbol{\Gamma}$, say $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}$, is

$$\left(\mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}} - \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}}\mathbf{J}^{-1}_{\boldsymbol{\Omega}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}}\right)^{-1}, \tag{31}$$

   that is, $\boldsymbol{\Omega}_0$ has no plugging-in effect on $\hat{\boldsymbol{\Gamma}}_0$.

Interestingly, the asymptotic variance of $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ can be written as a simple and transparent linear combination of avar$[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})]$ and avar$[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})]$. Explicit forms for these asymptotic variances can be computed from (31) and the formulas for the informa-

tion blocks given in the Supplement, as

$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})] =& \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Omega}, \\ \text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})] =& [\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T) - 2(\mathbf{I}_u \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T) \\ & + (\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)]^{-1}.\end{aligned} \tag{32}$$

The first equality can be obtained straightforwardly from $\mathbf{H}^T\mathbf{J}\mathbf{H}$, but the derivation of the second is quite involved – a detailed proof of (32) can be found in the Supplement, Section E. Also in the Supplement is a proof of how the theorem below follows from these equalities.

Theorem 6.1 *The asymptotic variance of* $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ *can be written as*

$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})] = & (\mathbf{I}_p \otimes \boldsymbol{\Gamma})\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})](\mathbf{I}_p \otimes \boldsymbol{\Gamma}^T) \\ & + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})](\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T).\end{aligned} \tag{33}$$

This representation can be made even more transparent if we recognize the following facts. Note that if $\boldsymbol{\Gamma}$ is known, then $\boldsymbol{\Gamma}\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is just $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$, the maximum likelihood estimate of $\boldsymbol{\beta}$. Similarly, for the second term in (33) we have

$$\begin{aligned}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})](\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T) &= \text{avar}[\sqrt{n}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})] \\ &= \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}\boldsymbol{\eta})].\end{aligned}$$

However, $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}\boldsymbol{\eta}$ is simply the maximum likelihood estimator of $\boldsymbol{\beta}$ when $\boldsymbol{\eta}$ is known. Hence we have:

Corollary 6.1 *The asymptotic variance of* $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ *has the following decomposition:*

$$\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})] = \text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] + \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})].$$

Intuitively, the asymptotic variance of $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ comprises those of $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ and $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})$; the role played by $\mathbf{Q}_{\boldsymbol{\Gamma}}$ is to orthogonalize these random vectors so that their contributions to the net asymptotic variance are additive.

Finally, to provide some insight on situations in which our estimator can be particularly effective, we compare $\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})]$ and $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$, the asymptotic variance of the usual MLE, in a relatively simple setting. Let $p = 1$, $\boldsymbol{\Omega} = \sigma^2 \mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = \sigma_0^2 \mathbf{I}_{r-u}$. In this case it can be shown that

$$\{\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}})\}^{-1/2}\{\boldsymbol{\Sigma}/\sigma_X^2\}\{\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}})\}^{-1/2} = \mathbf{I}_r + \frac{(\sigma_0^2 - \sigma^2)^2}{\sigma_X^2 \sigma^2 \|\boldsymbol{\beta}\|^2}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T \tag{34}$$

where we have used σ_X^2 in place of $\boldsymbol{\Sigma}_{\mathbf{X}}$ to emphasize that $p = 1$. This result indicates that the difference between our estimator and the standard MLE decreases when the signal ($\|\boldsymbol{\beta}\|$ or σ_X^2) increases, and increases when the variability (σ^2 or σ_0^2) increases. Equation (34) says also that the two approaches are equally efficient asymptotically when $\sigma^2 = \sigma_0^2$, a fact that is supported by the simulation results in Section 7. In full generality, (34) suggests that our estimator will provide the most gains in efficiency when the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be constructed from eigenspaces of $\boldsymbol{\Sigma}$ with relatively small eigenvalues (cf. Proposition 2.3). In particular, the size of u seems less important than the relative sizes of these eigenvalues, provided $u < r$.

7 Comparing Two Normal Means: simulation and data analysis results

We use the classic setting of comparing the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of two multivariate normal populations to illustrate the potential benefits of envelope models, and to verify our asymptotic calculations. In terms of model (1), the two-means comparison can be represented by taking $\boldsymbol{\alpha} = \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\mathbf{X} \in \{0, 1\}$. Since $p = 1$, we again use σ_X^2 in

place of $\mathbf{\Sigma_X}$ when describing various results.

In our simulations we tracked both small sample bias and variability. Since no appreciable bias was detected, Section 7.1 reports only variability comparisons, summarized using versions of the generalized standard deviation ratio

$$T = \{\mathrm{tr}(\mathbf{\Delta}_{\mathrm{em}}^{-1/2}\mathbf{\Delta}_{\mathrm{fm}}\mathbf{\Delta}_{\mathrm{em}}^{-1/2})/r\}^{1/2}, \tag{35}$$

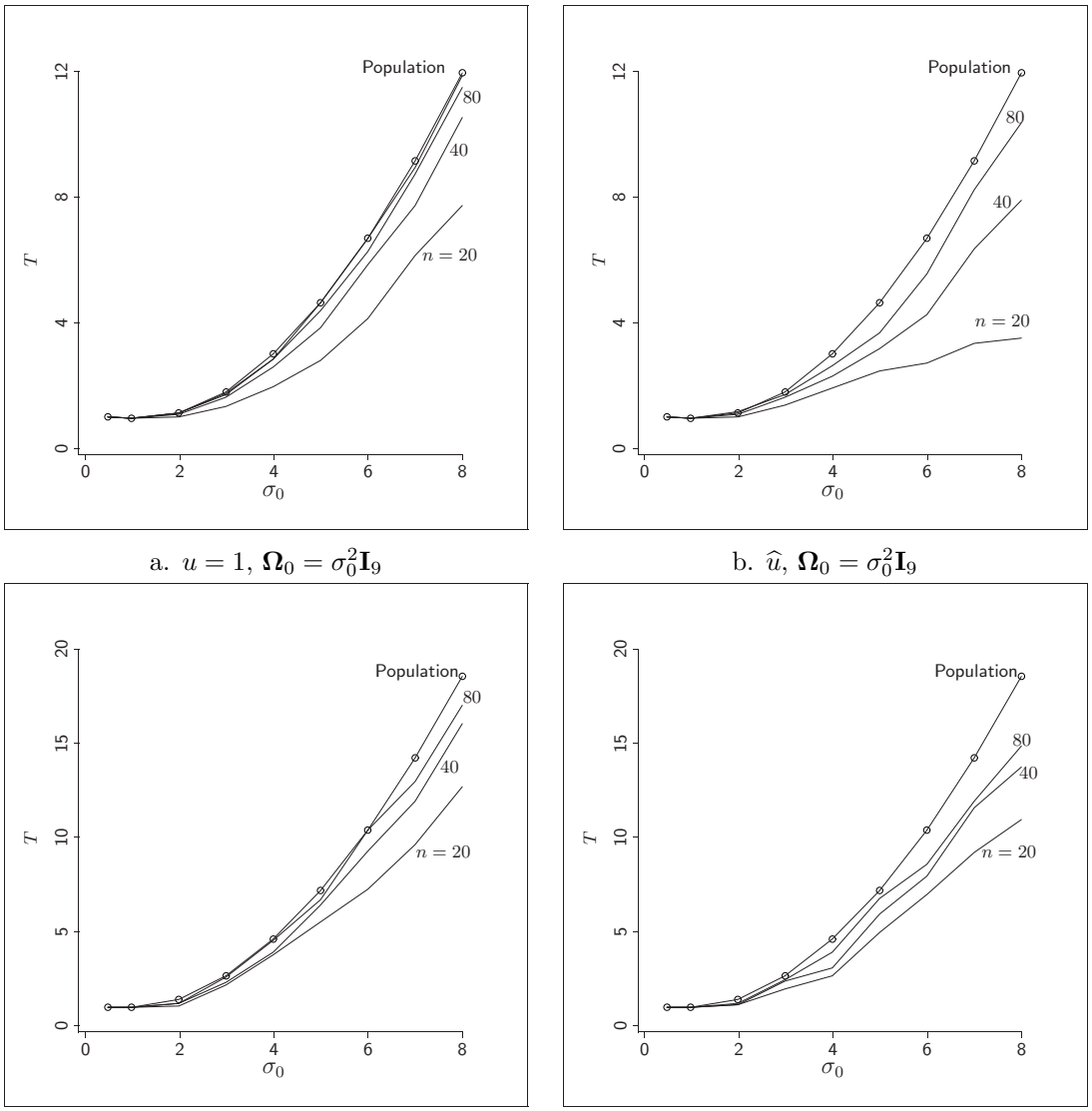
where $\mathbf{\Delta}_{\mathrm{fm}}$ represents the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$, the estimator of $\boldsymbol{\beta}$ from the full model (1), and $\mathbf{\Delta}_{\mathrm{em}}$ represents the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, the estimator of $\boldsymbol{\beta}$ from the envelope model (10) (for consistency of notation we use $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ instead of $\widehat{\boldsymbol{\beta}}$ to denote the envelope estimator in this section). T^2 can be interpreted as the average variance $\mathrm{E}(\boldsymbol{\ell}^T\mathbf{\Delta}_{\mathrm{fm}}\boldsymbol{\ell})$, where the average is computed over all $\boldsymbol{\ell} \in \mathbb{R}^r$ subject to the constraint that $\boldsymbol{\ell}^T\mathbf{\Delta}_{\mathrm{em}}\boldsymbol{\ell} = 1$. Values of $T > 1$ indicate that the envelope model (10) produces smaller standard deviations on average than the full model.

7.1 Simulation results

All results reported here were based on 200 replications from simulation models with $n/2$ observations per population, $r = 10$, $\boldsymbol{\beta}^T = (\sqrt{10}, \ldots, \sqrt{10})$, $u = 1$ and variance $\mathbf{\Sigma} = \sigma^2\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. Two versions of T were used. In the first, T_{pop}, we set $\mathbf{\Delta}_{\mathrm{em}} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}})$ (see (27)) and $\mathbf{\Delta}_{\mathrm{fm}} = \mathbf{\Sigma}/\sigma_X^2$, with all parameters at the values used in the simulations. It follows immediately from (34) that

$$T_{\mathrm{pop}}^2 = 1 + (1 - r^{-1})\frac{(\sigma^2 - \sigma_0^2)^2}{\|\boldsymbol{\beta}\|^2\sigma_X^2\sigma^2}.$$

In the second version, T_n, we set $\mathbf{\Delta}_{\mathrm{em}}$ and $\mathbf{\Delta}_{\mathrm{fm}}$ to be the sample covariance matrices of the 200 replications of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$. If our asymptotic calculations are correct, then for a sufficiently large n we should have $T_{\mathrm{pop}} \approx T_n$.



a. $u = 1$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_9$

b. $\widehat{u}$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_9$

c. $u = 1$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{A}^T \mathbf{A}/9$

d. $\widehat{u}$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{A}^T \mathbf{A}/9$

Figure 1: Simulation results for comparing the means of two multivariate normal populations.

The simulated data underlying Figure 1a were drawn using $\sigma^2 = 1$ and $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{I}_9$. We used the true value $u = 1$ when forming the estimate $\widehat{\boldsymbol{\beta}}_{\text{em}}$ based on (10). The upper curve, identified by the open circles, is a plot of T_{pop} for various values of σ_0. The other curves correspond to T_n for four samples sizes. As n increases T_n evidently approaches T_{pop} from below, with T_{80} being quite close to T_{pop}. The unlabeled curve that lies between T_{pop} and T_{80} was obtained with $n = 160$. The results in Figure 1a show that estimates from the envelope model can be much more efficient than the usual full-model estimates. They also support our previous conclusion that there is little difference between the methods when $\sigma \approx \sigma_0$.

Figure 1b was constructed as Figure 1a, except that, for the T_n curves, u was estimated as follows for each of the 200 simulated data sets. The hypothesis $u = u_0$ can be tested by using the likelihood ratio statistic $\Lambda(u_0) = 2(\widehat{L}_{\text{fm}} - \widehat{L}^{(u_0)})$, where $\widehat{L}_{\text{fm}}$ denotes the maximum value of the log likelihood for the full model, and $\widehat{L}^{(u_0)}$ the maximum value of the log likelihood for (10). Following standard likelihood theory, under the null hypothesis $\Lambda(u_0)$ is distributed asymptotically as a chi-squared random variable with $p(r - u_0)$ degrees of freedom. We employed the statistic $\Lambda(u_0)$ in a sequential scheme to choose u: Using a common test level of 0.01 and starting with $u_0 = 0$, we chose the estimate $\widehat{u}$ of u as the first hypothesized value that was not rejected. The results in Figure 1b show as expected that estimating u increases the variability of $\widehat{\boldsymbol{\beta}}_{\text{em}}$, but substantial gains are still possible for modest sample sizes. The drop for $n = 20$ is due mainly to the tendency of the likelihood ratio test to reject too frequently for small samples. The bounding dimension u could also be selected using an information criterion like AIC or BIC. Our intent here is to demonstrate only that reasonable inference on u is possible, without recommending a particular method.

Figures 1c and 1d were constructed as Figures 1a and 1b, except that $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{I}_9$ was replaced by $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{A}^T\mathbf{A}/9$, where $\mathbf{A} \in \mathbb{R}^{9\times 9}$ was generated once as a matrix of

standard normal variates. The range of the y-axis in Figures 1c and 1d is nearly twice that for Figures 1a and 1b, suggesting that correlation improves the performance of $\widehat{\boldsymbol{\beta}}_{\rm em}$ relative to $\widehat{\boldsymbol{\beta}}_{\rm fm}$.

7.2 Data analysis

We applied the proposed methodology to a number of data sets from the literature and found an advantage in most of them, suggesting that envelope models may have wide applicability. We present two brief illustrations in this section, one with a real but modest gain for envelope models and one with a dramatic gain.

In a sample of 172 New Zealand mussels, 61 were found to have pea crabs living within the shell and 111 were free of pea crabs. We compared the means of these two populations on $r = 6$ response variables, the logarithms of shell height, shell width, shell length, shell mass, muscle mass and viscera mass. Bartlett's test statistic for equality of covariance matrices has the value 27.8 on 21 degrees of freedom, so the assumption of equal covariance matrices seems reasonable. The p-values for the likelihood ratio tests of $u = 1$ and $u = 2$ were 0.024 and 0.18, suggesting that either of these values might be appropriate. Letting $\widehat{T}$ denote the estimate of T by using the plug-in method, we found that $\widehat{T} = 4.7$ for $u = 1$ and $\widehat{T} = 2.9$ for $u = 2$. In either case, it seems that the estimate of the mean difference from model (10) is notably less variable than the full-model estimate. Even with $u = 2$ these results indicate that it would take a sample about $2.9^2 = 8.41$ times as large for the efficiency of $\widehat{\boldsymbol{\beta}}_{\rm fm}$ to equal that of $\widehat{\boldsymbol{\beta}}_{\rm em}$ with the present sample size. The $\widehat{T}$ summary reflects the ratio of standard errors over all linear combinations of the coefficients. The standard error ratios are more modest when considering only individual coefficients, the individual standard errors for the full-model estimates ranging between 1.18 and 1.05 times the respective standard errors for the envelope estimates. For the largest of these, the envelope estimates achieve a reduction

equivalent to full-model estimates with a roughly 40 percent increase in sample size. We expect that this would be judged worthwhile in most analyses.

The second data set is the infrared reflectance example described in the Introduction. We chose this data set because the marginal response correlations are high, ranging between 0.9118 and 0.9991. This is the kind of situation in which the proposed methodology might give massive gains over a full-model analysis. The likelihood ratio test statistic for the hypothesis $u = 1$ has the value 1.09 on 5 degrees of freedom for a p-value of 0.95. With $u = 1$, $\widehat{T} = 219.2$. The standard deviation ratios for the individual mean differences were described in the Introduction.

To confirm the results for the infrared reflectance data we constructed a simulation model using all of the estimates from the original data as the population values. The population standard deviation ratio for this simulation scenario is $T_{\text{pop}} = 219.2$, which is the same as the plug-in estimate from the original data. We then constructed estimates based on 24 low protein observations and 26 high protein observations from the simulation model, repeating the process 200 times. This gave $T_n = 221.6$ and average plug-in estimate $\widehat{T} = 240.4$, which seems to support the results of the original analysis. To see if the high response correlations might introduce a notable small sample bias in $\widehat{\boldsymbol{\beta}}_{\text{em}}$ we computed element-wise $(\text{ave}(\widehat{\boldsymbol{\beta}}_{\text{em}})-\boldsymbol{\beta})/\boldsymbol{\beta}$, where $\text{ave}(\widehat{\boldsymbol{\beta}}_{\text{em}})$ denotes the replication average of $\widehat{\boldsymbol{\beta}}_{\text{em}}$. These six ratios ranged between -0.018 and 0.011. The same calculations using the 200 replications of $\widehat{\boldsymbol{\beta}}_{\text{fm}}$ produced six ratios ranging between -0.122 and 0.175.

The fit of model (10) to the original reflectance data gave $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2$, where $\widehat{\boldsymbol{\Sigma}}_1$ has rank 1 with non-zero eigenvalue 7.88 and $\widehat{\boldsymbol{\Sigma}}_2$ has rank 5 with eigenvalues $6,516.61$, 208.29, 20.08, 0.42 and 0.27. Evidently, the proposed method offers truly substantial gains in this example because the collinearity in $\boldsymbol{\Sigma}$ is quite large, and because $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is inferred to lie in an eigenspace of $\boldsymbol{\Sigma}$ with a relatively small eigenvalue.

8 Extensions and Relationships with Other Theory and Methods

In the previous sections we focused on one way in which the notion of enveloping can be employed; namely, creating a parsimonious, alternative parameterization for the multivariate linear model. However, envelopes can be used in other ways and in other contexts to allow more control over parameterizations and to develop methodology affording substantial gains in efficiency. We expect enveloping to have considerable potential in multivariate analysis: whenever we are dealing with a random vector $\mathbf{U}$ and an associated covariance matrix $\mathbf{\Lambda}$, we can consider a parsimonious parameterization of the latter in reference to the former. Mathematically, the essence of enveloping is to find the smallest reducing subspace of $\mathbf{\Lambda}$ to which $\mathbf{U}$ belong almost surely. In this section, we offer conjectures about a number of multivariate analysis contexts that share this form – the discussion is largely at the population level.

8.1 Reduced rank envelope models

Maximum likelihood estimation under model (1) does not require the coefficient matrix $\boldsymbol{\beta}$ to be of full rank $\min(r, p)$. Similarly, the envelope models introduced in the previous sections permit the rank of $\boldsymbol{\beta}$ to be less than $\min(r, p)$. In some regressions it may be useful to fit explicitly an envelope model with a specified rank d for $\boldsymbol{\beta}$. This, in effect, combines envelope models with models for multivariate reduced rank regression (Anderson, 1951; Izenman, 1975; Davies and Tso, 1982; Bura and Cook, 2003).

Recall from (10) that the mean function for the envelope model is $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\alpha}+\mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X}$, where $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$. If we restrict $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$ to have rank $d < \min(r, p)$, then $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ must have rank d and thus can be factored as $\boldsymbol{\eta} = \boldsymbol{\gamma}\boldsymbol{\phi}$, where $\boldsymbol{\gamma} \in \mathbb{R}^{u\times d}$ is a semi-orthogonal matrix and $\boldsymbol{\phi} \in \mathbb{R}^{d\times p}$ is unconstrained.

This gives a reduced rank version of envelope model (10):

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\gamma}\boldsymbol{\phi}\mathbf{X} + \boldsymbol{\varepsilon} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{36}$$

As before $\boldsymbol{\Gamma}$ is not identified, but $\text{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is identified and estimable. Similarly, $\text{span}(\boldsymbol{\gamma}) \in \mathbb{G}^{u\times d}$ and $\boldsymbol{\phi}$ are identified and estimable. Like the envelope version of model (1), this model has the potential for substantial gains in efficiency relative to the usual multivariate reduced rank model. Maximum likelihood and other methods of estimation for this model are currently under study.

It may be clear that we do not view reduced rank and envelope models as direct competitors, since combining them leads to a new model (36) which is more versatile than either one alone, and allows for more control over dimensionality. Similarly, many other methods for reducing dimensionality, like factor analysis, variable selection, and coefficient penalization (Yuan, Ekici, Lu and Monteiro, 2007), could be extended for use with envelope models.

8.2 Discriminant analysis

Consider classifying a new observation $\mathbf{y}$ on a feature vector $\mathbf{Y} \in \mathbb{R}^r$ into one of two normal populations C_1 and C_2, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common covariance matrix $\boldsymbol{\Sigma}$. Assuming equal prior probabilities, the optimal population rule, which is the same as Fisher's linear discriminant (Seber, 1984, p. 331), is to classify $\mathbf{y}$ as arising from C_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{y} > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2.$$

Letting $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$, $u \le r$, denote a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$, it follows from Corollary 2.1 that $\boldsymbol{\Sigma}^{-1}$ is of the form $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T$. The

optimal population rule expressed in terms of $\mathcal{E}_{\mathbf{\Sigma}}(\text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$ is to classify into C_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^T\mathbf{y} > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2.$$

Estimates of u, $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ can be found using the methods discussed in previous sections, specifying $\mathbf{Y}$ as the response vector. When $u \ll p$ or the eigenvalues of $\mathbf{\Omega}$ are substantially larger than those of $\mathbf{\Omega}_0$, we expect misclassification rates for this rule to be significantly lower than those for the standard rule. In cases where $u = 1$, $\mathcal{E}_{\mathbf{\Sigma}}(\text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) = \text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and, assuming that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Gamma} > 0$, the rule simplifies to $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{y} > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Extension to multiple populations with common covariance matrix seems straightforward conceptually.

Principal components have long been considered for dimension reduction prior to discriminant analysis. The first two methods discussed by Jolliffe (2002, Section 9.1) reduce $\mathbf{Y}$ by using the first few principal components from either the intra-population covariance matrix or the marginal covariance of $\mathbf{Y}$ computed without regard to population membership. Neither method is entirely satisfactory because there is no guarantee that the first few principal components will be the "best" for discrimination. The envelope approach proposed here has the potential to achieve what has long been attempted through principal component methodology.

8.3 Principal components

There are numerous ways to motivate the use of principal components for the reduction of a multivariate vector $\mathbf{Y} \in \mathbb{R}^r$ (Jolliffe, 2002). In this section we describe how an envelope construction might aid us in understanding a foundation for principal components based on latent variables (Tipping and Bishop, 1999).

Again consider model (1), only now $\mathbf{X} \in \mathbb{R}^p$ is an unobserved vector of latent variables, standardized to have mean 0 and variance $\mathbf{I}_p$, and $\boldsymbol{\beta}$ is assumed to have rank

$p < r$. The latent vector represents extrinsic variation in $\mathbf{Y}$, while the error $\boldsymbol{\varepsilon}$ represents intrinsic variation. The goal is to reduce the dimension of $\mathbf{Y}$ accounting for its extrinsic variation. Under this model it can be shown that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$ (Cook, 2007), and thus $\mathbf{R} = \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$ is the reduction we would like to estimate. Any full rank linear transformation $\mathbf{A}$ of $\mathbf{R}$ results in an equivalent reduction; $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{R}$ if and only if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{A}\mathbf{R}$, so it is sufficient to estimate $\mathcal{S} = \text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})$. Additionally, $\mathcal{S}$ is minimal; if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T \mathbf{Y}$ then $\mathcal{S} \subseteq \text{span}(\mathbf{B})$. Note that here we focus on estimating $\mathcal{S}$, though additional considerations may be necessary to translate knowledge about $\mathcal{S}$ into actions, depending on the application context.

Since $\mathbf{X}$ is not observed, only the marginal distribution of $\mathbf{Y}$ is available for the purpose of estimating $\mathcal{S}$. Following Tipping and Bishop (1999) we assume that $\mathbf{X}$ is normally distributed, and thus $\mathbf{Y}$ is normal with mean $\boldsymbol{\alpha}$ and variance $\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Sigma} + \boldsymbol{\beta}\boldsymbol{\beta}^T$. The maximum likelihood estimator of $\boldsymbol{\alpha}$ is just the sample mean of $\mathbf{Y}$, but $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are confounded and cannot be separated without additional structure. Tipping and Bishop (1999) assumed isotropic errors, i.e. $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_r$, and it follows from their results that the maximum likelihood estimator of $\mathcal{S} = \text{span}(\boldsymbol{\beta})$ is the span of the first p eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$, the sample version of $\boldsymbol{\Sigma}_{\mathbf{Y}}$. Consequently, $\mathbf{R}$ is estimated by the first p principal components of the marginal variance of $\mathbf{Y}$ when the errors are isotropic.

The assumption of isotropic errors is limiting relative to the range of applications in which it may be desirable to reduce multivariate observations. In the envelope parameterization of model (10), $\mathcal{S} = \text{span}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})$ and $\mathbf{Y}$ is normally distributed with mean $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Gamma}(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T$. The coefficients $\boldsymbol{\eta}$ are not identified since they are confounded with $\boldsymbol{\Omega}$, so it is still not possible to estimate $\mathcal{S}$. However, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\eta}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}^T \mathbf{Y}$ implies that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\Gamma}^T \mathbf{Y}$, so $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ provides an upper bound on the space of interest; $\mathcal{S} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. With isotropic errors, we have $\mathcal{S} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. We conjecture that, with a sufficiently large intrinsic signal $\boldsymbol{\eta}$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be estimated from the marginal of $\mathbf{Y}$. The

envelope model (10) with $\mathbf{X}$ as a latent vector would then allow estimation of the upper bound $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, which may be helpful in some applications and could provide insights onto the usefulness of principal components under a general error structure.

8.4 Envelopes in the predictor space

Recall from the discussion in Section 3 that the envelope model expressed by (10) has the greatest potential for improvement in regressions with many responses (r) and relatively few predictors (p). The novel parametrization we propose has nothing to offer when $r \leq p$ and $d = \dim(\mathcal{B}) = r$. This is the case, for example, in univariate linear regression ($r = 1$). Nevertheless, it may still be possible to achieve efficiency gains by using an envelope construction *in the predictor space*.

Assuming that $\mathbf{X}$ is random, the population coefficient matrix $\boldsymbol{\beta}$ can be represented as $\boldsymbol{\beta}^T = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\text{cov}(\mathbf{X}, \mathbf{Y})$, where $\mathbf{\Sigma}_{\mathbf{X}} = \text{var}(\mathbf{X})$ is the marginal variance of $\mathbf{X}$. Let $\mathcal{C} = \text{span}(\text{cov}(\mathbf{X}, \mathbf{Y})) \subseteq \mathbb{R}^p$ and let $\boldsymbol{\chi}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C})$, the $\mathbf{\Sigma}_{\mathbf{X}}$-envelope of $\mathcal{C}$. Then the coefficient matrix can be written as $\boldsymbol{\beta}^T = \mathbf{P}_{\boldsymbol{\chi}(\mathbf{\Sigma}_{\mathbf{X}})}\boldsymbol{\beta}^T$. This suggests that we estimate $\boldsymbol{\beta}^T$ by projecting the usual maximum likelihood estimate onto an estimate of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C})$, using the sample version of $\mathbf{\Sigma}_{\mathbf{X}}$ for the inner product. We would again expect notable efficiency gains if $\dim(\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C})) < p$ and we can find a good way to estimate $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C})$. One method of estimating $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C})$ that permits $n < p$ is described in Section 8.7.

8.5 Simultaneous envelopes

There is also the possibility of combining predictor space envelopes $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{C}) \subseteq \mathbb{R}^p$ with response space envelopes $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) \subseteq \mathbb{R}^r$ in a single multivariate regression. The predictor space envelopes of Section 8.4 rely on the identity $\boldsymbol{\beta}^T = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\text{cov}(\mathbf{X}, \mathbf{Y})$, which connects the coefficient matrix with population moment matrices. The corresponding expression

for the envelope model (10) follows from (2); $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\text{cov}(\mathbf{X}, \mathbf{Y})\mathbf{P}_{\boldsymbol{\Sigma}_1}$, where $\mathbf{P}_{\boldsymbol{\Sigma}_1}$ is still the projection onto $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. It now follows from Section 8.4 that

$$\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\text{cov}(\mathbf{X}, \mathbf{Y})\mathbf{P}_{\boldsymbol{\Sigma}_1} = \boldsymbol{\beta}^T\mathbf{P}_{\boldsymbol{\Sigma}_1} = \mathbf{P}_{\boldsymbol{\chi}(\boldsymbol{\Sigma}_{\mathbf{X}})}\boldsymbol{\beta}^T\mathbf{P}_{\boldsymbol{\Sigma}_1}.$$

This may serve as a conceptual starting point for the development of methods based on enveloping in both the predictor and response spaces.

8.6 Sufficient dimension reduction

There are various methods for reducing the dimension of a random predictor $\mathbf{X} \in \mathbb{R}^p$ in a regression with univariate response $Y \in \mathbb{R}^1$. Among them, *sufficient dimension reduction* methods estimate the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ (Cook, 1994, 1998), defined as the intersection of all subspaces $\mathcal{S}$ with the property that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Since the conditional distributions of $Y|\mathbf{X}$ and $Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ are identical, we can substitute $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ for $\mathbf{X}$ without loss of information on the regression.

Cook (2007) proposed that estimation of $\mathcal{S}_{Y|\mathbf{X}}$ be based on modeling the conditional distribution of $\mathbf{X}|Y$: Suppose that

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\varepsilon}, \tag{37}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\mathbf{f}_y \in \mathbb{R}^r$ is a known user-specified function of y, and $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}|Y}$. This is a multivariate linear model like (1), with the predictor vector $\mathbf{X}$ taking on the role of the response, and $\mathbf{f}_y$ taking the role of the predictor. However, in pursuing our sufficient dimension reduction, we have no particular interest in the coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$. Instead, interest lies in the central subspace $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\mathcal{B}$ (Cook, 2007), where still $\mathcal{B} = \text{span}(\boldsymbol{\beta})$. We can now use $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$ to parameterize (37), leading to an envelope model with the same form as

(10), or perhaps a reduced rank envelope model like (36). Because of the importance of $\mathcal{S}_{Y|\mathbf{X}}$ we might instead consider parameterizing in terms of the $\mathbf{\Sigma}_{\mathbf{X}|Y}$-envelope of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}|Y}}(\mathcal{S}_{Y|\mathbf{X}})$. In view of Proposition 3.1, however, these and several other envelopes are equal and thus lead to the same parameterization.

These considerations allow us a better understanding of the PFC model proposed by Cook (2007, eq. 13) without reference to envelopes:

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{f}_y + \boldsymbol{\varepsilon} \\ \mathbf{\Sigma}_{\mathbf{X}|Y} &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T. \end{aligned}$$

If we take $\mathbf{\Gamma}$ to be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$, then this is the envelope version of (37). Since $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$, the model only allows estimation of an upper bound on $\mathcal{S}_{Y|\mathbf{X}}$. To estimate the central subspace itself it is necessary to use a reduced rank envelope model, except in special cases where $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$.

8.7 Seeded reductions when $n < p$

In addition to the model-based approaches proposed by Cook (2007), there are numerous moment-based methods for estimating the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ without using models. Under various conditions, many of these approaches exploit population identities of the form $\mathcal{S}(\boldsymbol{\nu}) = \mathbf{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$, where $\mathbf{\Sigma}_{\mathbf{X}} = \text{var}(\mathbf{X})$ as defined previously, and $\boldsymbol{\nu}$ is a method-specific *seed* matrix (Cook, Li and Chiaromonte, 2007) that can be estimated from the sample moments of $(\mathbf{Y}, \mathbf{X})$ without inverting the sample version of $\mathbf{\Sigma}_{\mathbf{X}}$. For example, the least square seed (which corresponds to the multivariate linear model; see Section 8.4) is $\boldsymbol{\nu} = \text{cov}(\mathbf{X}, \mathbf{Y})$, and the seed for sliced inverse regression (Li, 1991) is $\boldsymbol{\nu} = \text{var}(\text{E}(\mathbf{X}|\mathbf{Y}))$. Of course, when n is sufficiently large, $\mathcal{S}_{Y|\mathbf{X}}$ can simply be estimated from the spectral structure of the sample version of $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\nu}$.

By using the $\mathbf{\Sigma_X}$-envelope of $\mathcal{S}_{Y|\mathbf{X}}$, Cook, Li and Chiaromonte (2007) developed a method of estimating $\mathcal{S}_{Y|\mathbf{X}}$ that does not require $n > p$. Their method is based on estimating $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{S}_{Y|\mathbf{X}})$ as the span of the sample version of $\mathbf{R}_u = (\boldsymbol{\nu}, \mathbf{\Sigma_X}\boldsymbol{\nu}, \ldots, \mathbf{\Sigma}_{\mathbf{X}}^{u-1}\boldsymbol{\nu})$, where u is also estimated. A basis for $\mathcal{S}_{Y|\mathbf{X}}$ is then estimated by projecting an estimate of $\boldsymbol{\nu}$ onto the space spanned by the estimate of $\mathbf{R}_u$, using the sample version of $\mathbf{\Sigma_X}$ to define the inner product. The method is equivalent to partial least squares (Helland, 1990) for univariate linear regression when the seed is $\boldsymbol{\nu} = \mathrm{cov}(\mathbf{X}, Y) \in \mathbb{R}^p$.

We conjecture that envelopes can be used in a variety of settings to develop estimation methods that allow $p > n$ and that generally have the potential to yield substantial gains in efficiency relative to standard methods, even when $p \ll n$. The particular method proposed by Cook, Li and Chiaromonte (2007) is a first step along these lines, but we expect that more efficient methods for estimating $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{S}_{Y|\mathbf{X}})$ are possible.

8.8 Functional data analysis

Although the envelope models at the core of this article concern multivariate linear regression, in which $\mathbf{Y}$, and possibly $\mathbf{X}$, are finite-dimensional random vectors, there is no fundamental difficulty in extending our approach to the case where $\mathbf{Y}$ and $\mathbf{X}$ are *random functions*; with an appropriate generalization, the ideas we presented are applicable to functional data analysis. The purpose of this subsection is to demonstrate this possibility by sketching such a generalization under somewhat strong simplifying assumptions. A fuller and more careful generalization will be considered in a future study. This generalization is significant because parsimony is even more important for functional data analysis.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $([0,1], \mathcal{G}, \lambda)$ the measure space where $\mathcal{G}$ is the class of Borel sets in $[0,1]$ and λ the Lebesgue measure. Next, let $L_2(\Omega, P)$ be the class of all random variables on Ω that are square integrable with respect to P, and

$L_2([0,1],\lambda)$ the class of functions defined on $[0,1]$ that are square integrable with respect to λ. Suppose $\varepsilon : \Omega \times [0,1] \to \mathbb{R}$ and $X : \Omega \times [0,1] \to \mathbb{R}$ are mappings such that, for each $t \in [0,1]$, $\varepsilon(\cdot,t)$ and $X(\cdot,t)$ are members of $L_2(\Omega,P)$. Thus $t \mapsto \varepsilon(\cdot,t)$ (or $t \mapsto X(\cdot,t)$) is a random function from $[0,1]$ to $L_2(\Omega,P)$, instead of a random vector from $\{1,\ldots,p\}$ (or $\{1,\ldots,r\}$) to $L_2(\Omega,P)$, as in the multivariate regression model (1). For simplicity, we assume both ε and X to be zero-mean functions; that is

$$\int_\Omega \varepsilon(\omega,t)P(d\omega) = 0, \quad \int_\Omega X(\omega,t)P(d\omega) = 0$$

for each $t \in [0,1]$.

Let $\kappa : [0,1] \times [0,1] \to \mathbb{R}$ be a bivariate kernel function and define

$$U(\omega,t) = \int_0^1 X(\omega,s)\kappa(s,t)\lambda(ds).$$

Assume the kernel κ is such that, for each $t \in [0,1]$, $U(\cdot,t)$ belongs to $L_2(\Omega,P)$. Define the functional linear regression model as $Y = U + \varepsilon$. This is a functional version of (1), except for ignoring the intercept – which has no bearing on this generalization.

Now, let $\Sigma : [0,1] \times [0,1] \to \mathbb{R}$ and $\Lambda : [0,1] \times [0,1] \to \mathbb{R}$ be the bivariate functions

$$\Sigma(s,t) = \int_\Omega \varepsilon(\omega,s)\varepsilon(\omega,t)P(d\omega), \quad \Lambda(s,t) = \int_\Omega U(\omega,s)U(\omega,t)P(d\omega).$$

For each $f \in L_2([0,1],\lambda)$, let $T_\Sigma(f)$ and $T_\Lambda(f)$ be the functions

$$t \mapsto \int_0^1 f(s)\Sigma(s,t)\lambda(ds), \quad t \mapsto \int_0^1 f(s)\Lambda(s,t)\lambda(ds).$$

Then, under mild conditions on Σ and Λ, T_Σ and T_Λ are bounded linear operators from $L_2([0,1],\lambda)$ to $L_2([0,1],\lambda)$.

Indicating with $\mathcal{B}$ the closure of the linear subspace $\text{span}\{T_\Lambda(f) : f \in L_2([0,1],\lambda)\}$,

the random function U belongs to $\mathcal{B}$ P-almost surely. We can then define the Σ-envelope of $\mathcal{B}$, say $\mathcal{E}_\Sigma(\mathcal{B})$, as the smallest reducing subspace of the linear operator T_Σ that contains the subspace $\mathcal{S}$. Furthermore, if we assume Σ to be such that T_Σ is a compact operator, then its spectral decomposition is very similar to that of $\boldsymbol{\Sigma}$ in model (1), with the eigenvectors of $\boldsymbol{\Sigma}$ replaced by the eigenfunctions of T_Σ. Essentially all the results we developed for the $\boldsymbol{\Sigma}$-envelope in the previous sections can be extended to this functional linear regression setting.

Restricting the size of the Σ-envelope is not only a means of parsimoniously modeling the variance operator T_Σ, but can also be used to constrain the mean function U. For example, if we assume that the Σ-envelope is finite-dimensional, then we have in effect constrained the mean to be a linear combination of a finite number of functions in $L_2([0,1],\lambda)$. Such a constraint can be useful when the sample size n is relatively small compared to the number of observations on each subject.

9 Conclusions

The results we presented in this article reveal a crucial property of the classical multivariate linear regression model (1); namely, that if the column space of the regression parameter $\boldsymbol{\beta}$ lies within a reducing subspace of the error covariance matrix $\boldsymbol{\Sigma}$, then far fewer parameters are needed to specify the likelihood. To express this parsimonious parameterization, we introduced the $\boldsymbol{\Sigma}$-envelope of $\boldsymbol{\beta}$, defined as the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\text{span}(\boldsymbol{\beta})$. The reparameterized likelihood can be maximized explicitly with the $\boldsymbol{\Sigma}$-envelope fixed, and maximization with respect to the latter can be performed numerically using Grassmann-manifold optimization. As we demonstrated analytically and on real and simulated data examples, this approach can bring dramatic improvements in accuracy relative to the traditional multivariate linear regression estimator.

We also argued that the notion of enveloping extends well beyond the parsimonious parameterization of the classical multivariate linear model. Obviously, any multivariate model that can be posed as a special case of (1) (e.g. a MANOVA model; Johnson and Wichern, 2007, Chapter 6) can be modified through a $\mathbf{\Sigma}$-envelope parameterization. Moreover, parameterizations based on error covariance envelopes could be devised for models that stem from generalizations of (1) – e.g. multivariate generalized linear models with non-Gaussian responses depending on the predictor through a multivariate non-linear link function (Fahrmeir and Tutz, 1994), or the functional linear models discussed in Section 8.8. Finally, reaching past linear models, we showed (Section 8) that enveloping can serve as a means to reinterpret, connect, and improve efficiency for a broad range of multivariate statistical techniques.

Acknowledgement

We would like to thank the editors, an associate editor, and three referees for their thorough and insightful reviews of several versions of the manuscript, which lead to significant improvements in both content and presentation. We also thank Inge Helland and Zhihua Su for their comments on previous versions. Zhihua reproduced all of the numerical calculations using independent code. Research for this article was supported in part NSF grants DMS-0704098 (R. Dennis Cook) and DMS-0704621 (Bing Li and Francesca Chiaromonte).

References

Aldrich, J. (2005). Fisher and regression. *Statist. Sci.* **20**, 401–417.

Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327–351.

Bura, E. and Cook, R. D. (2003). Rank estimation in reduced-rank regression. *J. Mult. Anal.* **87**, 159–176.

Christensen, R. (2001). *Advanced Linear Modeling.* Springer, New York.

Conway, J. (1990). *A Course in Functional Analysis.* Second Edition. Springer, New York.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences*, pp. 18-25. American Statistical Association, Alexandria, VA.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22**, 1–26.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

Davies, P. T. and Tso, M. K.-S. (1982). Procedures for reduced rank regression. *Applied Statist.* **31**, 244–255.

Edelman, A., Tomás, A. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–353.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer, New York.

Faraway, J. and Reed, M. P. (2007). Statistics for digital human motion modeling and ergonomics. *Technometrics* **49**, 277–290.

Helland, I. S. (1990). On the structure of partial least squares regression. *Scandinavian J. Statist.* **17**, 97-114.

Henderson, H. V., and Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian J. Statist.* **7**, 65–81.

Hoffman, K., Zyriax, B. C., Boeing, H., and Windler, E. (2004). A dietary pattern derived to explain biomarker variation in strongly associated with the risk of coronary heart disease. *Am. J. Clin. Nutr.* **80**, 633-40.

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *J. Mult. Anal.* **5**, 248–264.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Sixth Edition. Pearson Prentice Hall.

Jolliffe, I. T. (2002). *Principal Component Analysis, 2nd ed.* Springer, New York.

Leek, J. T., and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, 1724–1735.

Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316–327.

Li, K. -C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Am. Statist. Assoc.* **98**, 99–109.

Liu, X., Srivastava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 662–666.

Reinsel, G. C. and Velu, P. (1998). *Multivariate Reduced Rank Regression, Theory and Applications.* Lecture Notes in Statistics 136. Springer, New York.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *J. Am. Statist. Assoc.* **81**, 142–149.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components. *J. Royal Statist. Soc., Ser. B* **61**, 611-622.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Royal Statist. Soc., Ser. B* **69**, 329–346.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist.* **38**, 1092–1109.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
Email: dennis@stat.umn.edu.

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.
Emails: bing@stat.psu.edu and chiaro@stat.psu.edu.